Doctoral Dissertation

Korean Genome Analysis

Sungwon Jeon

Department of Biomedical Engineering

Ulsan National Institute of Science and Technology

2021

Korean Genome Analysis

Sungwon Jeon

Department of Biomedical Engineering

Ulsan National Institute of Science and Technology

Korean Genome Analysis

A dissertation submitted to
Ulsan National Institute of Science and Technology
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Sungwon Jeon

05.12.2021 of submission

Approved by

Advisor

Jong Bhak

Korean Genome Analysis

Sungwon Jeon

This certifies that the dissertation of Sungwon Jeon is approved.

05.12.2021 of submission

Signature

Advisor: Jong Bhak

Signature

Semin Lee

Signature

Seung Woo Cho

Signature

Dougu Nam

Signature

Jeongeun Kim

Abstract

A personal genome can now be analyzed very efficiently at a low cost by using high-throughput whole-genome sequencing technologies, along with fast and accurate computing methods such as machine learning. Very large-scale population genomics studies that extensively investigate the whole ethnic groups are now plausible. Ethnic genome data are crucial resources for mapping population-specific genomic patterns and identifying diseases- and phenotypes-associated variants for use in healthcare. Genomics data can also provide insight into the population histories of both ancient and modern ethnic groups. Although there have been several Korean personal and populational genomic studies in the past 20 years, very large-scale Korean population genomic data with matched phenotypes have not been made available. Further, the study of the origin and composition of the Korean population based on whole-genome study and multiomics data, have not been thoroughly studied. In this Ph.D. dissertation, I present my analysis of Korean genomes. In the first chapter, I present a variome set from the first phase of Genome Korea (Korea1K, 1094 Korean genomes) which is a subproject of the Korean Genome Project (KGP) and its usefulness. The Korea1K variome analysis showed that the Korean population is genetically highly homogenous compared to other East Asians. The Korea1K variome and its matched clinical traits data illustrated the significant advantage of using whole-genome sequences for genome-wide association studies, by identifying nine more significant candidate alleles than previously reported. As a reference variome panel for the population genomics, the Korea1K panel showed better imputation accuracy for Koreans than the commonly used 1,000 genome project panel (1KGP) of the United Kingdom. In the second chapter, I describe my investigation into the origin and genomic composition of the Korean population using 88 Korean whole-genome data accompanied by 208 worldwide and 115 ancient genomes from the various eras and spatial spectrums. This extensive comparative analysis suggested that the current genomic composition of Koreans may have been established through rapid admixture events between ancient southern Chinese associated with Bronze-Iron age Southeast Asians and existing Northern Asians around and inside of the Korean peninsula. I also speculated that the admixing trend initially occurred mainly outside the Korean peninsula, followed by continuous spread and localization within the Korean peninsula, which is consistent with the general admixture trend of East Asians in the Bronze and Iron ages that occurred about 4,500 years ago. The genomics composition of more than 70% of modern Koreans' is thought to be derived from the recent population expansion and admixture events from the South. In the third chapter, I introduce the first systematically produced Korean Genome Project portal and its open API system, which allows the variant frequencies and association results of Korea1K to be efficiently accessible. In conclusion, I present a large-scale Korean genome analysis, thereby showing the usefulness of constructing the population variome set. The Korean variome analysis, in combination with worldwide modern and ancient genomic resources, can also be used to explain the origin of Koreans.

Contents

List of Figures

List of Tables

Nomenclature

Preface **1**

I. Introduction **3**

Genomics technology 3

Korean genome studies 4

II. Chapter 1. Korean population genome analysis **6**

2.1. Introduction 6

2.2. Methods 7

2.3. Results 13

2.4. Discussion 39

III. Chapter 2. Korean origin **40**

3.1. Introduction 40

3.2. Methods 41

3.3. Results and discussion 46

3.4. Conclusion 64

IV. Chapter 3. Korean Genome Project portal **66**

4.1. Introduction 66

4.2. Methods 67

4.3. The Korean Genome Project Portal 69

4.4. Discussion 73

V. Conclusion **74**

References **75**

Acknowledgments **90**

Appendix **92**

List of Figures

Fig. 1 Principal component analysis (PCA) plot using SNVs and Indels in Korea1K set

Fig. 2 Boxplot of variants quality normalized by depth based on allele frequency category and existence in dbSNP v.150 before and after batch effect filtering

Fig. 3 Variants statistics and discovery rate of the novel variants.

Fig. 4 Percentage of overlapped SNVs with KoVariome.

Fig. 5 Number of variants from variome databases based on allele frequencies.

Fig. 6 Variants distribution based on variant location and allele frequency in Korea1K.

Fig. 7 Fraction under selection based on variant type.

Fig. 8 Fraction under selection based on genes.

Fig. 9 Length distribution of indels.

Fig. 10 Length distribution of Indels in the coding region.

Fig. 11 Number of novel variants as a function of new unrelated individuals.

Fig. 12 Proportion of variants based on allele categories for A) PolyPhen and B) SIFT estimation.

Fig. 13 Mitochondrial haplogroup distribution in Korea1K.

Fig. 14 Chromosome Y haplogroup distribution in Korea1K.

Fig. 15 Comparison with other populations. Results of principal component analysis of Korea1K and the 1KGP set of (A) worldwide populations and (B) East Asian samples.

Fig. 16 *ADMIXTURE* plot for Korea1K and 1KGP East Asians.

Fig. 17 ClinVar variants which have more than 10% of allele frequency in the Korea1K.

Fig. 18 Drug response variants found in Korea1K. Blue indicates significantly different allele frequencies between the Korea1K dataset and the population from the Chi-square test.

Fig. 19 Length distribution of copy number variations.

Fig. 20 Copy number variations in Korea1K.

Fig. 21 Transposable element (TE) insertion frequency distribution in Korea1K.

Fig. 22 PCA plot using Transposable element (TE) insertion.

Fig. 23 Transposable element (TE) insertion frequency distribution of Korea1K and 1KGP populations.

Fig. 24 Significance of TE insertion allele frequency difference.

Fig. 23 Transposable element (TE) insertion frequency distribution of Korea1K and 1KGP populations.

Fig. 24 Significance of TE insertion allele frequency difference.

Fig. 25 HLA allele distribution in Korea1K.

Fig. 26 Comparison of HLA type frequency to the public database.

Fig. 27 Manhattan plot of the reported loci via a genome-wide association study.

Fig. 28 QQplots for the GWA tests of the 20 traits.

Fig. 29 QQplots for the GWA tests of the 20 traits.

Fig. 30 QQplots for the GWA tests of the 20 traits.

Fig. 31 QQplots for the GWA tests of the 19 traits.

Fig. 32 Minor allele frequency (MAF) of the most significant variant on the loci from GWA analysis.

Fig. 33 Imputation performance evaluation.

Fig. 34 FineSTRUCTURE analysis of 88 Koreans and 208 contemporary global individuals.

Fig. 35 Global inference of the genetic structures.

Fig. 36 Global inference of the genetic structures after filtration (*K*=2-14)

Fig. 37 Dendrogram correlation between the fineSTRUCTURE clade and ADMIXTRUE (*K*=2-14) results.

Fig. 38 Genetic clustering of the present-day populations

Fig. 39 A global outgroup *f3* statistics between the ancient and present-day populations

Fig. 40 Genetic association between the ancient and present-day populations.

Fig. 41 *D*-statistics with a form of *D*(Yoruba, Tianyuan, ancient, present-day)

Fig. 42 Genetic cluster of the ancient genomes analyzing with outgroup f3 statistics

Fig. 43 *D*-statistics with a form of *D*(Yoruba, Tianyuan; E_{si}, $EA_{a/b}$)

Fig. 44 Bronze and Iron Age gene flows making up the Korean

Fig. 45. D-statistics with a form of *D*(Yoruba, ancCS; ancient, present-day)

Fig. 46 Haplotype distribution in the Korean population

Fig. 47 Admixture tree model depicting the historical genetic makeup of Korean

Fig. 48. Four tested admixture tree models by qpgraph.

Fig. 49. Architecture of the Korean Genome Project portal.

Fig. 50. Homepage of the Korean Genome Project portal.

Fig. 51. Sample result of an API request.

Fig. 52. Sample gene page of the Korean Genome Project portal.

Fig. 53. Sample variant page of the Korean Genome Project portal.

List of Tables

Table 1 Variant count before and after removing batch effect.

Table 2 Number of Transposable element (TE) insertions before and after filtering.

Table 3 List of traits with index variants located in previously reported loci.

Table 4. Outgroup *f3*(Korean, present-day population, Yoruba) analysis

Table 5. Admixture *f3* statistics

Table 6. Estimation of admixture date of Koreans

Nomenclature

NGS, next-generation sequencing

KGP, Korean genome project

SNV, single-nucleotide variant

Indel, insertion, and deletion

CNV, copy number variation

TE, transposable element

HLA, human leukocyte antigen

PCA, principal component analysis

1KGP, the 1000 genome project

GWAS, genome-wide association study

EA, East Asia

SEA, Southeast Asia

KGPP, Korean Genome Project Portal

Preface

This dissertation is the results of my five-and-a-half years of a graduate course at the Department of Biomedical Engineering (BME), Ulsan National Institute of Science and Technology (UNIST) and Korean Genomics Center (KOGIC), UNIST, supervised by Professor Jong Bhak.

The main work in this dissertation is mainly based on the Ulsan 10,000 genome project (Korea10K), which aimed to sequence more than 10,000 Korean whole-genomes for future use as biomedical resources, as a part of the Genome Korea in Ulsan project and the Korean Genome Project (KGP). Participation in this project was voluntary. The participants of this project donated their blood and clinical information, and submitted a lifestyle survey result. As a welfare reward, we provided them with personal genome research reports that were obtained by analyzing their genomes as a welfare reward. My lab, KOGIC, has completed sequencing of 10,000 Korean whole-genomes in April 2021. Since Ulsan city is my hometown and the place from where I completed my entire education, it was a great honor for me to be involved in the Genome Korea in Ulsan project as a researcher and a citizen of Ulsan.

The first chapter describes my analyses on 1,094 Korean whole-genomes (Korea1K), performed in collaboration with KOGIC, Genome Research Foundation, and Clinomics members. Of the whole genomes, 1,007 samples were studied for the first two years (2016-2017) of the project. By developing an efficient variant-calling pipeline for the large-scale population genome data, I was able to generate the Korea1K variome. I also developed the Korean Genome Project Portal (KGPP) to share the Korea1K variome data efficiently via a web portal system; this is introduced in the third chapter. I believe that this kind of very large-scale Korean population genome data should be shared as openly and freely as possible to facilitate future genomic and clinical studies among Koreans and other populations with minimal regulation.

Another topic presented in this dissertation is an analysis of the Korean origin in the population using whole-genome resources. In 2017, my collaborators and I published the first Neolithic East Asian genome data (Devil's Gate Cave), which presented our findings related to the recent dual origin of the Korean population and the continuous genomic composition of East Asians. However, we did not have ancient Southeast Asian genomes at the time of publication, as a result of which we were unable to investigate Korean origin properly. Fortunately, due to advancing ancient genome sequencing technology, many ancient Southeast Asian genomes were published in 2018. Using the now available data on ancient genomes near the Korean peninsula and hundreds of modern human genomes, my colleagues and I were able to construct a much more detailed model of Korean origin. Our model showed conclusively that the Korean ethnic group has been formed almost entirely from the South in tens of thousands of years and through admixture with surrounding ethnic groups, with the most

significant events of this group formation occurring during the Bronze-Iron age explosions. I believe that this work has an important social and national significance as it may lead to a more scientific understanding of East Asians, especially the Chinese, Japanese, and Koreans, who are extremely close to each other ethnically. My work may also help lay the foundation for better understanding of each other among these population in the future, thereby leading to more peaceful and cooperative interactions among the major East Asian populations. The second chapter of this dissertation presents my investigation into the construction of the model of Korean origin.

The appendix consists of a list of abstracts of the research papers in which I am one of the authors. This dissertation somehow covers these papers in parts or as a whole.

Sungwon Jeon, Ulsan, Republic of Korea, 2021

I. Introduction

Genomics technology

A genome is defined as the entire genetic material and information of an organism, a population, and/or a species. In *Homo sapiens*, the genome consists of the protein coding and the non-coding regions of chromosomes. A genomic sequence is a series of nucleotides or bases (A, C, G, and T for DNA) in a specific order that presents complete genomic information relayed as a specific order of bases. The human genome contains over three billion bases (3 Gb) and more than 20,000 protein-coding genes[1]. The genome can be sequenced using genome sequencing techniques, such as Sanger sequencing, shotgun short read sequencing, next-generation sequencing (NGS), and next-next generation sequencing techniques such as nanopore sequencing.

Next-generation sequencing produces a massive number of short and long sequence reads. A sequenced individual genome is usually analyzed using two different approaches, depending on whether a reference genome is available. If the reference genome is not available, *de novo* reference genome assembly must be performed. The reads can be overlapped and extended to contigs through sequence assembly. Using other sequencing technologies like Hi-C, long-read sequencing, and 10X chromium, the contigs can be extended to scaffolds or even whole chromosomes[2]. The constructed reference genome represents the genomic structure of the organism or individuals and can be compared with that of organisms from the same species or close-distance species or long-distance species[3-5]. Comparative genomics presents genomic and evolutionary differences between individuals or organisms that can be used to understand the biological features of the genomes of organisms, which will thus facilitate the development of target biological molecules (biomarkers)[3].

On the other hand, if the reference genome is available, the sequencing reads can be mapped to the reference, and individual variants can be discovered by contrasting the sequencing reads against the reference. A set of the found variants from individuals can be defined as “Variome,” and the variome can be used to investigate diseases or phenotypes-associated variants using known genomic markers, or to examine the ancestry history of the sample’s population[6-8].

The history of a single human or a specific population can also be investigated by analyzing the variome. For example, the large-scale variome constructed in the 1000 genome project from 2,540 individuals across 29 worldwide populations has been used to impute genotypes as a reference panel, to find evolutionarily selected variants[9]. The Genome Aggregation Database(gnomAD), which contains a much larger set of variome constructed from 141,456 samples across multiple populations, was recently published and is freely available[10]. The gnomAD variome has been used to annotate frequencies of target genomic variants and to investigate population genomic selection pressures.

Another application of the mapping approach is in an analysis of ancient genomes. DNA samples from ancient specimens can be sequenced using developed ancient DNA sequencing library preparation technology. It could be possible to identify genome-wide genotypes in ancient genomes[11], which, in turn, enables researchers to investigate human migration, history, and specific regional migration and admixture patterns. Ancient genome analysis can also provide the signatures of positive selection of extinct species[12].

Korean genome studies

The Korean population is distributed worldwide (majorly in the Korean peninsula) and comprises around 100 million people. It is largely a very homogeneous population, with a few large-scale admixture events having occurred in the past, thus necessitating the genomics study of this population necessary. In 2008, the first Korean whole-genome sequence (SJK) was published[13]. The sequencing reads of SJK were mapped to the human reference, and its genotypes were identified. The subsequent comparison of SJK with four published human genomes and showed 3,782,072 differences (variants) between SJK and the human reference. The first Korean *de novo* reference genome was assembled using various types of genome sequencing libraries[14]. Cho *et al.* used short-read and mate-pair sequencing data from eight different DNA libraries and constructed primary contigs. The contigs were further assembled into scaffolds (KOREF_S) with optical mapping information, thereby generating a chromosome-level of *de novo* reference genome assembly. In particular, they aligned 40 Korean whole-genome sequencing data from the KoVariome database to KOREF_S and substituted the alleles of KOREF_S with the major allele of the Korean population (KOREF_C). They suggested that KOREF_S and KOREF_C might be useful in improving the alignment of East-Asian personal genomes for efficient variant-calling and large-scale population genome analysis.

The KoVariome database, presented in 2018, comprises an analysis of 50 Korean whole-genome sequences from the Korean Personal Genome Project (KPGP)[15]. Although the KPGP was the biggest large-scale of the Korean population genomic analysis at the time of its publication, it could not cover the variants with allele frequency less than 1% due to the small sample size. Further, the project did not include matched clinical information for association tests. Another study presented the KOVA database developed from 1,055 Koreans[16]. The authors investigated the characteristics of small variants and copy number variants (CNVs) in the Korean population. However, since their database was based on exome-sequencing that cannot cover the entire human genome, it is unable to fully elucidate variant profiles across the whole genome. With respect to the ancient Korean genome, Veronika *et al.* presented the first Neolithic East Asian genome, called Devil's Gate, which showed high genomic continuity between ancient and modern Koreans[11]. However, due to the lack of availability of ancient genome from Southeast Asia, this study could not thoroughly investigate the origin of Koreans, which is critical to model the origin of the Koreans and East Asians.

Although previous Korean genomics studies have tried to provide the genomic characteristics of Koreans, none have been large-scale (>1,000) Korean population genomic studies, and no study has investigated the origin of the Korean population. In this Ph.D. dissertation, I outline my research on the Korean genome analysis. In chapter 1, I present a large-scale Korean population genomic study, Korea1K. Korea1K showed Korean-specific variome patterns using 1,094 whole-genomes and investigated the genomic utility of the large-scale Korean variome set. In chapter 2, I present my analysis of the origin of the Korean population, conducted using ancient and modern human genomes, including 88 Korean whole-genomes. The results of this analysis indicated a complex admixture model of the Korean population derived from a recent population expansion and global and local admixture from the South to East Asia. In the chapter 3, as an application of Korea1K, I introduce the Korean Genome Project portal and its open application programming interface (API) system which provides the variome data and phenome association of Korea1K through a web application.

This doctoral dissertation is an addition based on the following papers that the author has already published.

Sungwon Jeon, *et al.* Korean Genome Project: 1094 Korean personal genomes with clinical information. *Science advances* 6.22 (2020)

Jungeun Kim, Sungwon Jeon, *et al.* "The origin and composition of Korean ethnicity analyzed by ancient and present-day genome sequences." *Genome biology and evolution* 12.5 (2020): 553-565.

II. Chapter 1. Korean population genome analysis

2.1. Introduction

The Korean population [estimated census population size close to 85 million (M)] has been thought to be highly homogeneous with few large-scale admixture events in the past[11, 17-19]. However, little formal scrutiny has been given to these claims. Several Korean whole genomes and exomes[16, 20] have been reported since the first Korean genome data (SJK) were published in 2008[13], including the first Korean reference genome sequence (KOREF_S)[14] and 40 unrelated individuals (KOREF_C) that formed the basis of KoVariome, the Korean genomic variation database[15]. Before the current study, at least 100 whole genomes of Korean individuals were available worldwide[20, 21]. However, although a global whole-genome project (the multiethnicity 1000 genome project) that aims to characterize global human genetic diversity contains over 2,500 genomes, including Chinese and Japanese, it does not include Korean samples yet[9].

There has also been an effort to generate ethnicity-specific reference genome sequences, and several human variomes have been generated to expand the coverage of human genome diversity, including the UK10K[22], the Genome of the Netherlands (GoNL) project[23], and the pan-African genome[24]. In 2015, the consequences of strong founder effects were demonstrated in the Icelandic population by sequencing 2,636 genomes[25]. In the Danish population study, 150 trios were used to de novo assemble a reference genome, and they provide detailed data on structural variations and many complex genomic regions, including the major histocompatibility complex and major regions of the Y chromosome[26]. In East Asia, the 1KJPN project yielded data on 1,070 Japanese genomes[27], and another recent dataset identified selection signatures in the Japanese population from 2,234 Japanese whole-genome data[28]. In contrast, the original KoVariome database contained only 50 Korean whole-genome sequences without clinical information at the time of publication[15], although its sample size has subsequently increased to >100 genomes. Despite these large genome sequencing projects in numerous populations, little biochemical and clinical data and limited information regarding genotype-phenotype association for the participants have been collected to characterize the population's health and disease states.

In this chapter, I introduce the Korea1K dataset comprising 1,094 Korean whole genomes of which 1,007 genomes were newly generated in combination with systematically acquired clinical and biochemical measurement information from the blood and urine of the participants. This Korea1K set represents the first-phase release of the Korean Genome Project (KGP). KGP is a joint project by the Personal Genome Project at Harvard Medical School, the National Center for Standard Reference Data of Korea, Clinomics Inc., and the Korean Genomics Center of Ulsan National Institute of Science and Technology (UNIST). These genomes have been sequenced to a high sequencing depth (~31× on average) using Illumina HiSeq X10, and I used these data to characterize single-nucleotide variants

(SNVs), indels, copy number variations (CNVs), transposable element (TE) insertion, and human leukocyte antigen (HLA) type in the Korean population and contrast the Korean data with similar data from other populations. The majority of the genomic data (984 samples) were from volunteers with clinical information on 79 quantitative traits that were measured at Ulsan University Hospital. To evaluate the practical utility of this large genomic dataset, I performed a genome-wide association study (GWAS) using the information of the 79 quantitative clinical traits. I also quantified the effectiveness of the dataset as a reference panel by analyzing 19 previously published Korean gastric cancer patient genomes[29].

2.2. Methods

2.2.1. Sample collection and sequencing

Informed consent was obtained from all individuals for their participation in the Korean Ulsan genome project, which comprises two subprojects. All clinical information was examined by the Ulsan University Hospital. In total, 696 samples were curated in the Ulsan University Hospital Biobank, from which samples were received thereafter. Further, 311 samples were collected by us. I downloaded data from 87 Korean samples from KoVariome[15], which collected volunteers from all across South Korea. Sample collection and sequencing was approved by the Institutional Review Board (IRB) of the Ulsan National Institute of Science and Technology (UNISTIRB-15-19-A and UNISTIRB-16-13-C). Genomic DNA was isolated from human blood samples, using the DNeasy Blood & Tissue kit (Qiagen, Germany) in accordance with the manufacturer's protocol. Genomic DNA from saliva samples was isolated using the GeneAll Exgene trademark clinic SV mini kit. Extracted DNA was quantified using the Quant-iT BR assay kit (Invitrogen). High–molecular weight genomic DNA was sheared using a Covaris S2 ultra sonicator system to obtain fragments of appropriate sizes. Libraries with short 350–base pair (bp) inserts for paired-end reads were prepared using the TruSeq Nano DNA sample prep kit in accordance with the manufacturer's protocol for Illumina-based sequencing. The products were quantified using the Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA), and raw data were generated using an Illumina HiSeq X10 platform (Illumina). Clusters were generated using paired-end 2 × 150-bp cycle sequencing reads via resequencing. Further image analysis and base calling were carried out using the Illumina real-time analysis program (https://sapac.illumina.com/informatics/sequencing-data-analysis.html) with default parameters following the manufacturer's instructions. The quality of the base in the read was checked by FastQC (ver. 0.11.5; www.bioinformatics.babraham.ac.uk/projects/fastqc/).

2.2.2. Variant calling

Adapter contamination was trimmed using Cutadapt (ver. 1.9.1)[30] with a forward adapter (′GATCGGAAGAGCACACGTCTGAACTCCAGTCAC′) and reverse adapter

('GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT') and with a minimum read length of 50 bp after trimming. Thereafter, trimmed reads were mapped to the hg38 reference, using BWA-MEM (ver. 0.7.16a) with the "-M" option[31]. Mapped BAM files were sorted by coordination, using Picard (ver. 2. 14. 0) with the Sortsam module. Duplicate reads were marked using Picard (ver. 2. 14. 0) with the MarkDuplicates module. The mapping quality was recalibrated using the BaseRecalibrator tool in the Genome Analysis Tool Kit (GATK) (ver. 3.7)[32]. gVCF files were generated by HaplotypeCaller in GATK[32] with "-stand_call_conf 30-ERC GVCF" option. SNVs and indels were jointly genotyped from the gVCF files by GenotypeGVCFs in GATK[32]. The called variants were annotated using Variants Effect Predictor (VEP) ver. 92[33], and the fraction was estimated under negative selection, using the script of Moon and Akey (https://github.com/moon-s/fraction-under-selection)[34]. For estimation of a fraction under selection pressure for each protein-coding gene, I selected genes that have more than 250 alternative allele count sums, since the small number of allele count may not produce proper site frequency spectrums. The following annotated variants were assigned to LoF mutations: "Frame_Shift_Del", "Frame_Shift_Ins", "In_Frame_Del", "In_Frame_Ins", "Nonsense_Mutation", "Nonstop_Mutation", or "Splice_Site".

2.2.3. Calling of copy number variations

Copy number variations (CNVs) were identified via CNVnator[35] with default parameters and a 100-bp bin size from 1,094 samples. Thereafter, I excluded 23 samples in accordance with the following criteria: the total number of CNV exceeds one standard deviation (SD) from the average count of CNVs per sample (average CNV count: 525, SD of CNV count: 129). Reliable CNVnator calls were filtered in accordance with the following criteria:

1) e-values (e-val1, e-val2, e-val3, and e-val4) are less than 10^{-5}

2) q0 <0.5 (q0 is the fraction of reads mapped with zero quality)

3) Gap and centromere regions from UCSC hg38 data were filtered out.

4) For deletion calls, only those with <0.75 of normalized read depth*(1+q0) were used.

5) If the bases in the called region contained more than 90% of the "N," the calls were filtered out.

6) Segmental duplication regions from UCSC hg38 data were filtered out.

After individual calling, the calls with >80% reciprocally overlapped regions from each individual were combined using the igraph package[36] in R. The start and end positions of the representative calls were assigned to the average of the locations from the combined calls. I annotated the gene symbol of the CNV calls with Ensembl database[37]. I then checked the overlap between the call set and 1KGP phase

3[38]. Only CNVs that showed more than 80% of the overlap with CNVs of 1KGP were used for further analysis. I additionally used Control-FREEC[39] with a window size 100 bp and a breakpoint threshold 0.6 to validate the common CNVs which contained protein-coding genes in Korea1K. I filtered out the common CNVs which have lower than 0.85 of the recovery rates of the CNV calls between the two callers.

2.2.4. Calling of TE insertions

TE insertions were identified in Korea1K samples using Mobile Element Locator Tool (MELT; ver. 2.1.4)[40], a tool to detect TE insertions in ALU, LINE1, and SVA elements using discordant read pairs to define potential TE sites and split reads to identify breakpoints and target site duplications[38]. I filtered out TE sites with <70% and >130% of average depth of 100bp flanking regions to control for variations at candidate TE sites. The allele frequency of TE insertions was calculated as the number of presented TE insertions normalized by the total number of alleles in the population[41]. I compared Korea1K TE insertions with 1KGP by the genomic position and TE types. Only TE insertions with frequencies of >5% and which overlapped with 1KGP call were used to PCA. Chi-square analysis was performed for each TE insertion and TE insertions with Q-values of <0.05 were determined to identify TE insertions differing significantly from those in the Korea1K dataset.

2.2.5. HLA typing

The HLA gene complex encodes MHC proteins responsible for antigen presentation. HLA typing was carried out using OptiType (ver. 1.3.1)[42], which predicts information regarding HLA class 1 alleles from WGS data. Reads were mapped to HLA reference sequences in the OptiType program using BWA[31] (ver. 0.7.15) and all unmapped reads were filtered out using SAMtools (ver. 1.6)[43]. Thereafter, I run OptiType's pipeline with default parameters. To compare frequencies from multiple populations, I downloaded an HLA allele frequency database from The Allele Frequency Net Database[44].

2.2.6. Batch effect removal

Each sample was labeled in accordance with its sequencing library preparation protocol, sequencing company, and the date of sending blood samples or libraries to the company. Twelve technical batches were identified. The batch effect was assessed via PCA, using EIGENSOFT (ver. 6.1.4)[45], using variants and samples in accordance with the following criteria:

For variants:

1) Biallelic SNVs with a MAF of ≥5%.

2) *P* values of the Hardy-Weinberg Equilibrium (HWE) test >0.05.

3) Genotype missing rate of <0.01.

Thereafter, filtered variants were pruned on the basis of linkage disequilibrium (LD), using PLINK[46] (ver. 1.9b) with "--indep-pairwise 200 4 0.1", leaving 101,326 SNVs. For individual selection, closely related individuals identified on the basis of an identity by descent (IBD), estimated in PLINK[46], were filtered. All pairs with an IBD value of >0.125 were extracted (corresponding to third-degree relatives) and clustered to a family group. Until no pairs of relatives remained, each family group was then reduced as follows:

1) The sample with the highest number of pairs in the family group was eliminated.

2) The sample with the highest missing calls among LD-pruned SNVs was eliminated if there are several samples with the same number of pairs in the family group.

To identify variants exhibiting the batch effect, I used logistic regression models for all variants as follows:

1) The variant was eliminated if it was a batch-specific variant compared to all other batches.

2) Each batch was paired with another, resulting in all possible combinations. The variant was eliminated if it was significant in any of the combinations.

In total, 6,348,049 variant positions were significantly associated with the technical batch ($P \leq 0.01$) and eliminated from the original set. I used the quality by depth (QD) value in a joint VCF file for plotting variants' quality distribution.

2.2.7. PCA and *ADMIXTURE* with the 1KGP genome data

The interpopulation genomic structure was evaluated by projecting the first two PCs determined via PCA of SNVs from Korea1K samples and 1KGP without closely related individuals. I selected and merged variants and from the Korea1K and 1KGP sets in accordance with the following criteria:

1) Biallelic SNVs with a MAF of ≥5%.

2) Biallelic SNVs with an HWE $P > 10^{-6}$.

3) Biallelic SNVs with a missing genotype rate of <0.01.

Extracted variants were LD pruned using "--indep 50 5 2" in PLINK[46], yielding 153,633 sites. PCA was carried out using the EIGENSOFT program[45]. *ADMIXTURE*[47] analysis was performed from $K = 2$ to $K = 14$ based on the same variants set as PCA. I plotted an ADMIXTRUE plot for $K = 3$, which showed the smallest cross-validation error rate across the Ks.

2.2.8. Mitochondrial and chromosome Y haplogroup analysis

Mitochondrial haplogroups were identified via Haplogrep (ver. 2.1.13)[48, 49], and the Yfitter tool (ver. 0.2)[50] was used to identify Y chromosome haplogroups. I prepared the input files for the Yfitter program

by converting hg38 coordination to hg19 coordination, using CrossMap[51] (ver. 0.2.7).

2.2.9. GWAS

For the GWAS, 823 individuals, 6,658,227 variants, and 79 traits were selected in accordance with the following criteria:

For individuals:

1) Individuals whose clinical traits were examined.

2) Individuals having no rare diseases.

3) Individuals having no kinship within the selected samples.

For variants:

4) SNVs and indels having a MAF of ≥1%.

5) SNVs and indels having an HWE $P > 10^{-6}$.

6) SNVs and indels having a missing genotype rate of <0.01.

The GWAS was performed exclusively using quantitative traits. GWA analysis was performed using linear regression under an additive genetic model. Age, age2, sex, body mass index (BMI), and the first 10 principal components were included as covariates. BMI was excluded from covariates in the GWAS for BMI itself and the degree of obesity. The genome-wide significance threshold was determined to be 7.51×10^{-9} through Bonferroni correction (0.05/6,658,227). The study-wide significance threshold was determined using the equation (0.05 / (the number of tested traits × the number of tested variants)). Variants were grouped into the loci with "--clump-p1 0.0000001 --clump-kb 1000 --clump-r2 0.1" options with PLINK (version 1.9)[46]. For each locus, previously reported variants associated with the trait of interest were examined in order from the most significant variant with the National Human Genome Research Institute (NHGRI) GWAS catalog[52] ($P \leq 5 \times 10^{-8}$, ver. 2018-12-07).

2.2.10. Imputation panel construction

To construct the Korea1K imputation reference panel, 1,059 healthy individuals that had no rare diseases with a total of 28,692,913 autosomal biallelic variants with a missing genotype call rate of <0.1 and minor allele count of >1 (not a singleton) were selected. The variants were phased into haplotype using SHAPEIT2 (version v2.r904)[53], and the Korea1K set was used to construct a rephased imputation panel using the 1KGP reference panel. I chose an alternative allele from Korea1K if the alternative allele of Korea1K and 1KGP is discordant during the merging step. To evaluate the imputation accuracy of the reference panels, I separately processed matched normal samples from previously published 19 unrelated Korean patients with gastric cancers obtained from National Center for Biotechnology

Information (NCBI; SRP014574 and SRA057772). The WGS data from 19 previously reported Korean individuals with gastric cancer were mapped to hg38 using BWA-MEM (ver. 0.7.15) with the "-M" option, and the SAM format was converted to the BAM format using SAMtools (ver. 1.4)[43]. The BAM files were sorted using SAMtools (ver. 1.4), and duplicated reads were marked using the MarkDuplicates module in Picard tools. Base realignment and recalibration of the base quality score were carried out using GATK (ver. 3.7)[32]. Variants from all samples were called using GATK HaplotypeCaller with the joint calling mode. For a test set, I extracted 1,302,490 variants that were present in Illumina Omni 2.5 chip from the 19 individuals and obtained 1,243,087 prephased SNVs using SHAPEIT2[53]. The prephased test set was imputed using the prepared reference panels by Minimac3 (ver. 2.0.1)[54]. Imputation accuracies were estimated using squared Pearson correlation coefficients (R^2) between the true genotypes and imputed genotype dosages.

2.3. Results

2.3.1. SNVs and Indels in Korea1K dataset

Whole-genome sequencing (WGS) data from 1,007 blood or saliva samples (984 samples with clinical and biochemical information) were generated with an average sequencing depth of 31× and pooled with sequencing data from an additional 87 blood or saliva samples (without clinical information) from the KoVariome database[15]. In total, 1,094 complete genomes, including 916 unrelated and healthy individuals, mostly from the Ulsan metropolitan region, were compared to the human genome reference (hg38). A total of 39.2 M SNVs and 7.6 M indels were called from the dataset (Table 1). I filtered out false-positive variants due to the sequencing batch effect (Figs. 1 and 2) and related individuals, yielding a set of variants containing 34 M SNVs and 4.8 M indels. I divided the variants into five categories based on their allele frequency in the Korean population (singleton: allele count = 1; doubleton: allele count = 2; rare: allele count of >2 and allele frequency of ≤0.01; common: allele frequency of >0.01 but ≤0.05; and very common: allele frequency of >0.05; Fig. 3A). Highlighting the power of the large dataset, approximately half of the variants that I identified were classified as singleton or doubleton (allele count of ≤2), and unexpectedly, more than 70% of them are not reported in dbSNP (v150)[55]. On the other hand, less than 20% of the variants were classified as very common (allele frequency of >0.05), with more than 94% of these variants previously reported in dbSNP (v150) (20). A total of 96.6% of the very common SNVs overlapped with KoVariome[15], compared to only 12.4% of rare SNVs (Fig. 4). The number of variants that have an allele frequency of >0.01 was similar to other non-African populations (1KGP non-African and 3.5KJPN), while there was a far higher number of variants, which have an allele frequency of ≤0.01 than KoVariome because of Korea1K's much larger sample size (Fig. 5).

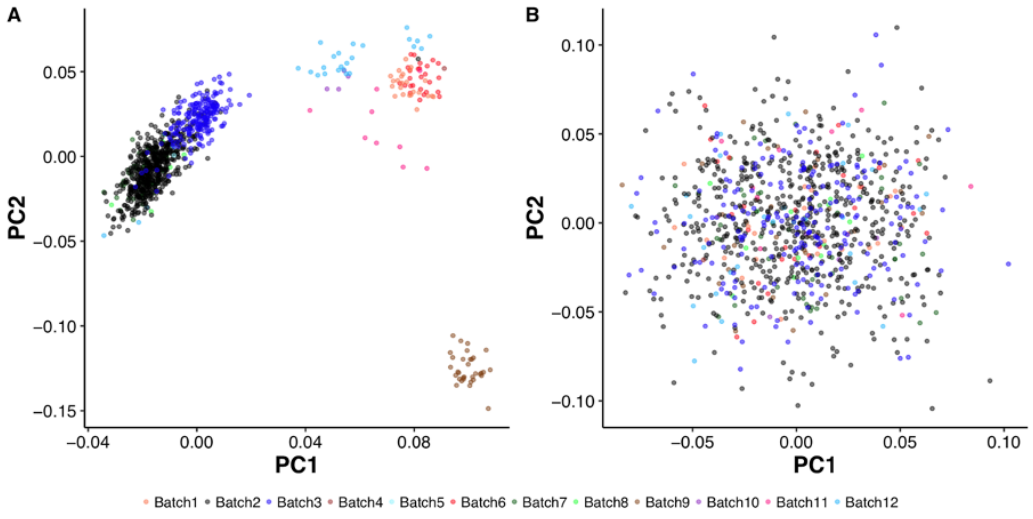


Fig. 1 Principal component analysis (PCA) plot using SNVs and Indels in Korea1K set. (A) before removing the batch effect (B) after removing the batch effect.

Table 1 Variant count before and after removing batch effect. Singleton: allele count =1; Doubleton: allele count =2; rare: allele count > 2 and allele frequency ≤ 0.01; common: allele frequency > 0.01 and allele frequency ≤ 0.05; very common: allele frequency > 0.05. The allele frequency category in this table was based on 1,094 individuals.

Variant type	Allele frequency category	Reported	Before removing batch effect	After removing batch effect	Remaining percentage
SNV	Very Common	Novel	237,739	68,290	28.72%
	Very Common	dbSNP	6,122,812	4,589,385	74.96%
	Common	Novel	536,623	410,651	76.53%
	Common	dbSNP	2,104,239	1,612,718	76.64%
	Rare	Novel	7,777,745	6,988,611	89.85%
	Rare	dbSNP	5,380,191	4,212,315	78.29%
	Doubleton	Novel	3,034,057	2,804,907	92.45%
	Doubleton	dbSNP	1,549,851	1,365,313	88.09%
	Singleton	Novel	8,787,498	8,729,585	99.34%
	Singleton	dbSNP	3,659,762	3,434,077	93.83%
	Total SNV		39,190,517	34,215,852	87.31%
Indel	Very Common	Novel	673,458	236,402	35.10%
	Very Common	dbSNP	1,478,967	850,680	57.52%
	Common	Novel	942,429	408,410	43.34%
	Common	dbSNP	543,954	280,808	51.62%
	Rare	Novel	1,407,307	900,776	64.01%
	Rare	dbSNP	563,066	383,256	68.07%
	Doubleton	Novel	440,245	362,171	82.27%
	Doubleton	dbSNP	107,841	88,174	81.76%
	Singleton	Novel	1,191,058	1,118,475	93.91%
	Singleton	dbSNP	207,223	180,358	87.04%
	Total Indel		7,555,548	4,809,510	63.66%
Total variants			46,746,065	39,025,362	83.48%

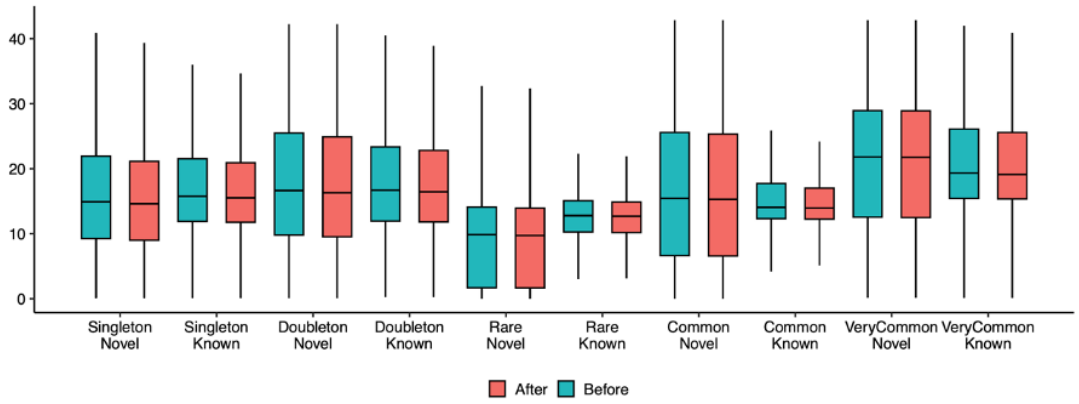


Fig. 2 Boxplot of variants quality normalized by depth based on allele frequency category and existence in dbSNP v.150 before and after batch effect filtering. (Singleton: allele count =1; Doubleton: allele count =2; rare: allele count > 2 and allele frequency ≤ 0.01; common: allele frequency > 0.01 and allele frequency ≤ 0.05; very common: allele frequency > 0.05)

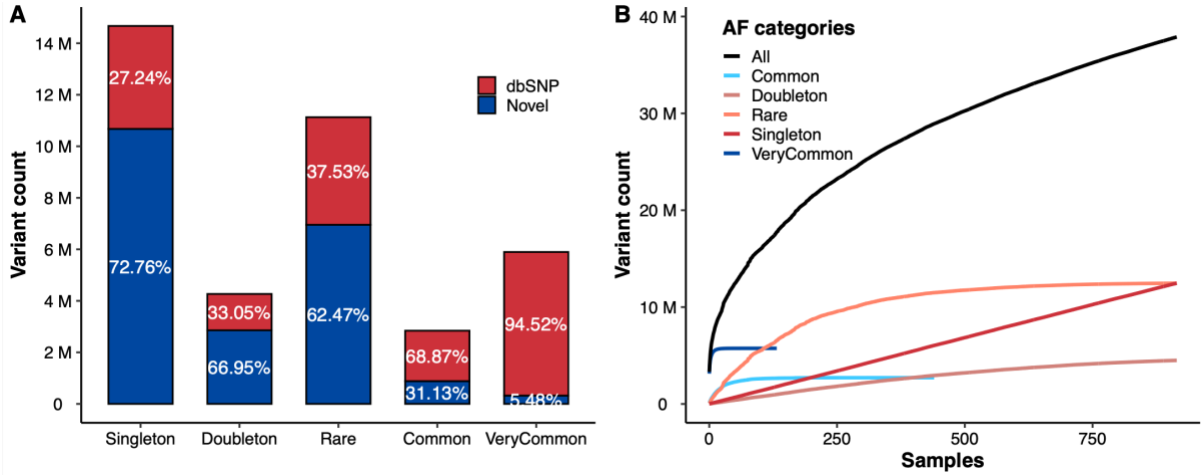


Fig. 3 Variants statistics and discovery rate of the novel variants. (A) Number of variants in the Korea1K dataset in all autosomal regions categorized based on allele frequencies. Singleton: allele count =1; Doubleton: allele count =2; rare: allele count > 2 and allele frequency ≤ 0.01; common: allele frequency > 0.01 and allele frequency ≤ 0.05; very common: allele frequency > 0.05. (B) The number of novel variants as a function of unrelated Korean genome samples.

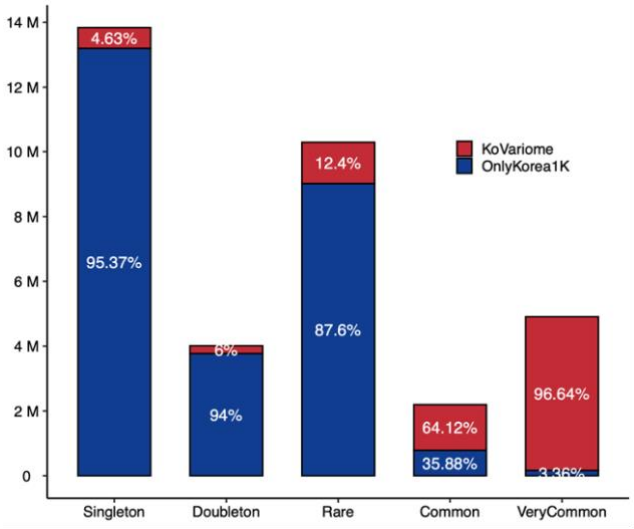


Fig. 4 Percentage of overlapped SNVs with KoVariome. Singleton: allele count =1; Doubleton: allele count =2; rare: allele count > 2 and allele frequency ≤ 0.01; common: allele frequency > 0.01 and allele frequency ≤ 0.05; very common: allele frequency > 0.05

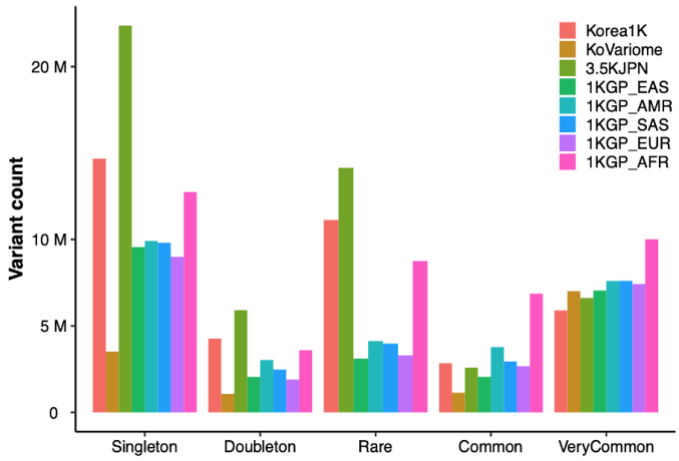


Fig. 5 Number of variants from variome databases based on allele frequencies. Singleton: allele count =1; Doubleton: allele count =2; rare: allele count > 2 and allele frequency ≤ 0.01; common: allele frequency > 0.01 and allele frequency ≤ 0.05; very common: allele frequency > 0.05. The color indicates variome database. Note that there are no variants that have allele frequency ≤ 0.01 and allele count >2 for KoVariome because of the small size of samples.

On the basis of the final set of variants, each individual showed on average ~4.42 M variants (3.58 M very common, 0.4 M common, 0.31 M rare, 0.46 M doubleton, and 0.85 M singleton variants), of which 8,928 and 918 were nonsynonymous and loss of function (LoF), respectively. Next, I classified each variant into 1 of 19 different variant classes (i.e., intergenic and intronic) based on its functional impact and location in the genome (Fig. 6). LoF variants (nonsense, nonstop, splicing site, and indel variants) in the Korea1K set had a higher ratio of rare, doubleton, or singleton variants than other regional classes, indicating the effect of purifying selection on these variants. In addition, the allele (site) frequency spectrum of unrelated individuals was used to estimate the fraction under selection pressure in different genomic regions[34]. I confirmed that LoF variants had the highest fraction of sites under negative selection (Fig. 7). I applied the same comparative analysis to the entire gene set and found that 16 genes showed high purifying selection pressure, which was even stronger than the selection for nonsynonymous variants across the genome. Four genes showed negative values suggestive of positive selection pressure (Fig. 8). Regarding indels, the Korea1K set displayed more deletions (2,573,411) than insertions (2,155,644), possibly resulting from skewed variant calling (Fig. 9). Indels in protein-coding regions displayed higher peaks among in-frame indels based on their length, indicating purifying selection (Fig. 10)[56].

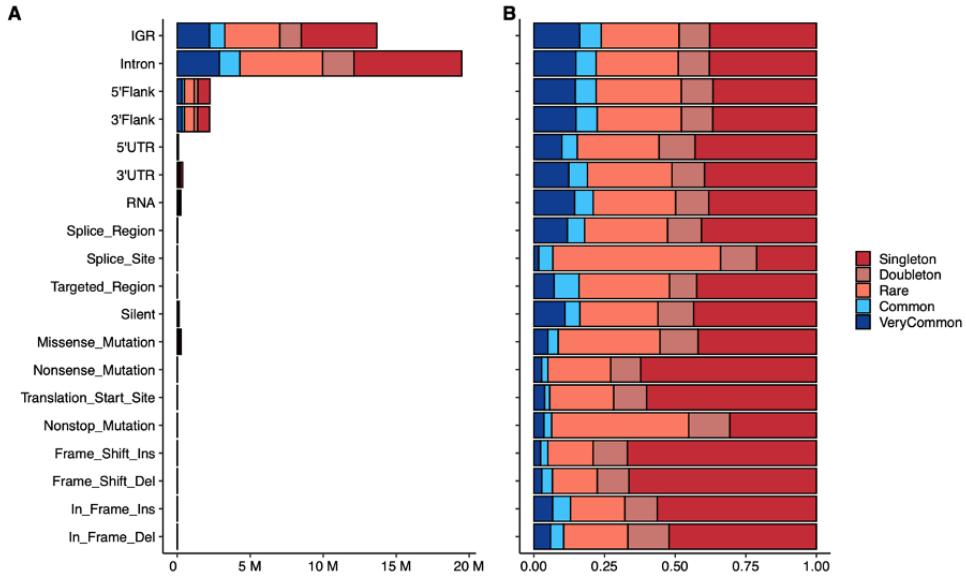


Fig. 6 Variants distribution based on variant location and allele frequency in Korea1K. (A) Variants counts, and (B) proportions of the number of variants based on allele frequency categories. IGR: inter-genic region except for 5' and 3' Flank variants; UTR: untranslated region. Singleton: allele count =1; Doubleton: allele count =2; rare: allele count > 2 and allele frequency ≤ 0.01; common: allele frequency > 0.01 and allele frequency ≤ 0.05; very common: allele frequency > 0.05

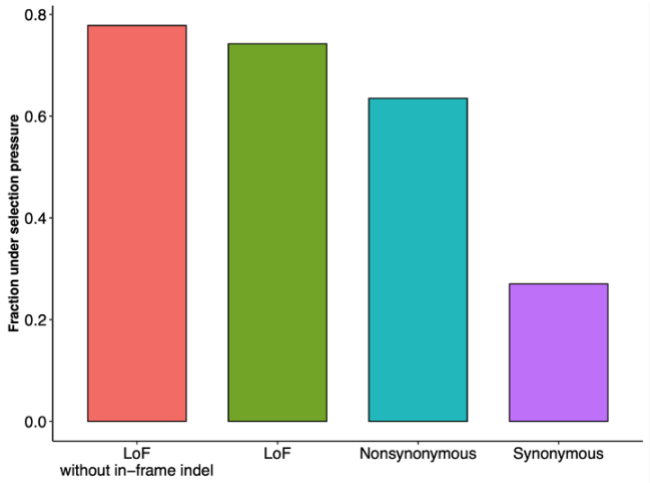


Fig. 7 Fraction under selection based on variant type. LoF indicates loss-of-function.

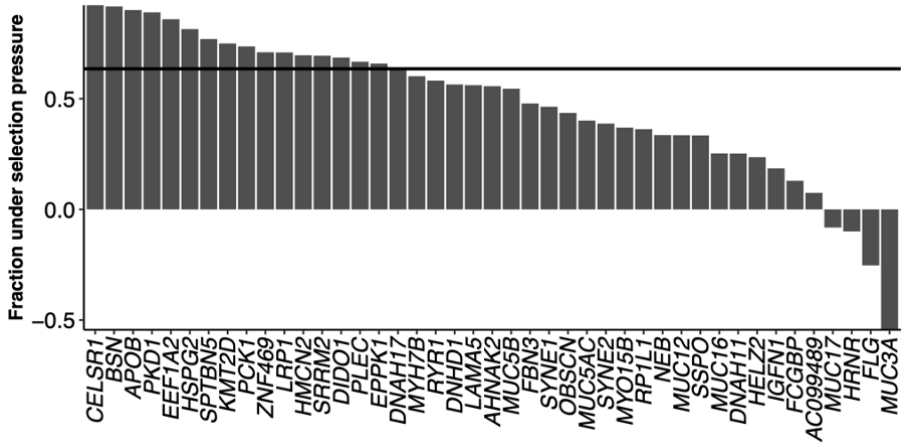


Fig. 8 Fraction under selection based on genes. The horizontal line indicates the fraction under the selection pressure of nonsynonymous variants.

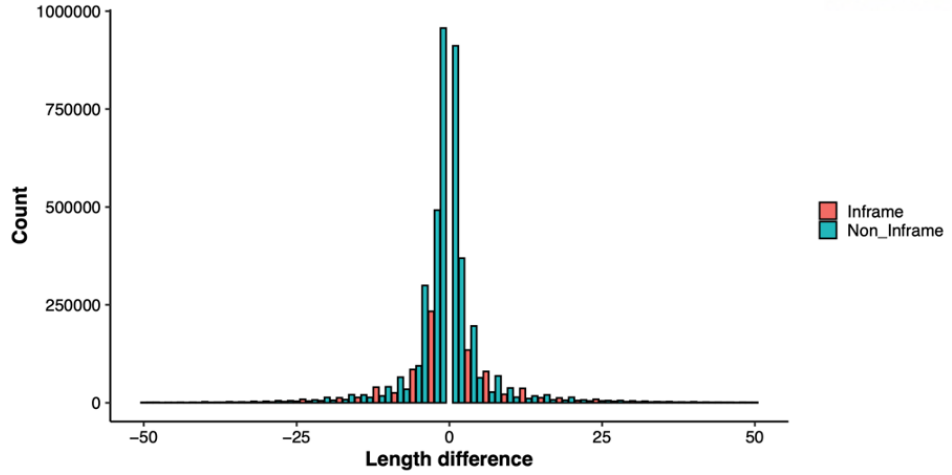


Fig. 9 Length distribution of Indels.

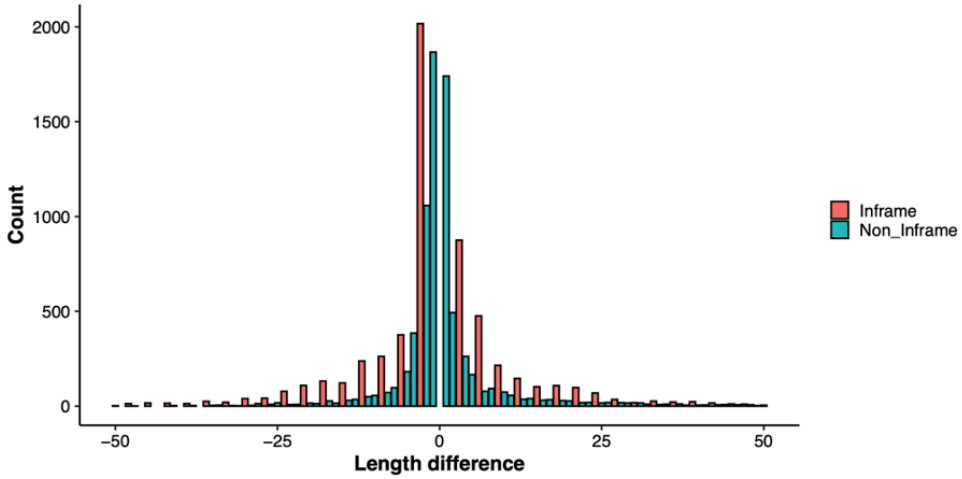


Fig. 10 Length distribution of Indels in the coding region.

The discovery rate of newly observed variants from unrelated individual genomes is a method for quantifying genomic diversity in a given population[57]. The pattern of newly observed unshared variants of unrelated Korean genomes was investigated using the five allele frequency categories (Fig. 3B). The discovery rate of very common (allele frequency of >0.05) variants saturated after 132 samples (14.4%), while the rate of singleton and doubleton variants was still increasing after analyzing all 916 healthy unrelated samples. When I compared the count of newly observed variants in unrelated individuals against previously published KoVariome, unexpectedly, Korea1K showed a slightly higher rate of novel variant discovery than KoVariome (Fig 11; Korea1K, 101,866; KoVariome, 48,051 for 50th individual). This increase might have been caused by implementing newer versions of the variant calling pipeline and the human genome reference. As expected, this also confirms that more sequenced genomes are needed to sufficiently cover very rare variants in the Korean population.

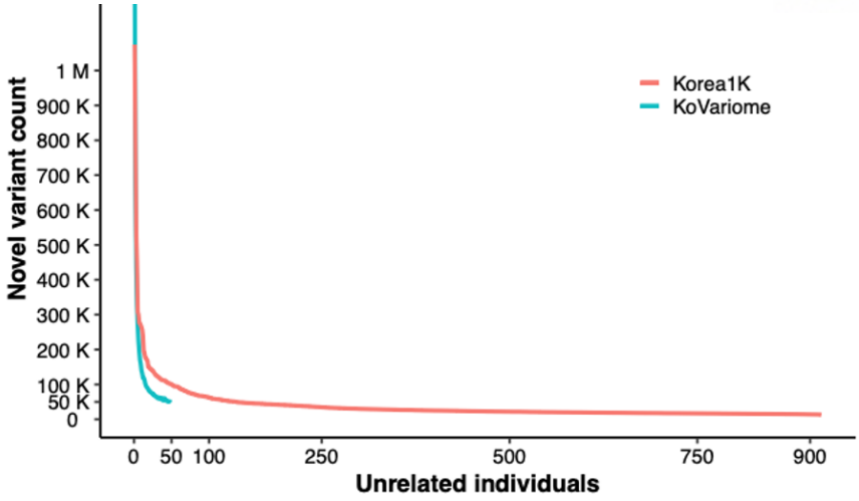


Fig. 11 Number of novel variants as a function of new unrelated individuals.

The Korea1K set contained 266,081 nonsynonymous SNVs. Among them, 118,417 and 117,414 were categorized as protein damaging by PolyPhen[58] (possibly damaging, 46,116; probably damaging, 72,301) and SIFT [59] (deleterious, 117,414), respectively. In total, 87,671 variants were predicted as protein damaging by both programs, and their allele frequency is skewed toward rare frequencies, while benign or tolerated variants are skewed toward common frequencies, again indicating purifying selection (Fig 12). When mitochondrial and chromosomal Y haplogroups among the Korean individuals (Figs 13 and 14) were investigated, the common types identified were D (34.19%), B (13.89%), and M (13.80%) mitochondrial and O (73.49%), C (16.9%), and N (6.58%) chromosomal Y[60, 61]. The O male haplogroup is widely distributed in East Asia and Southeast Asia, while the C haplogroup is prominently distributed in East Asia and Northeast Asia[60]. I also identified other fairly common mitochondrial haplogroup types (A, G, and F) in East Asia[62].

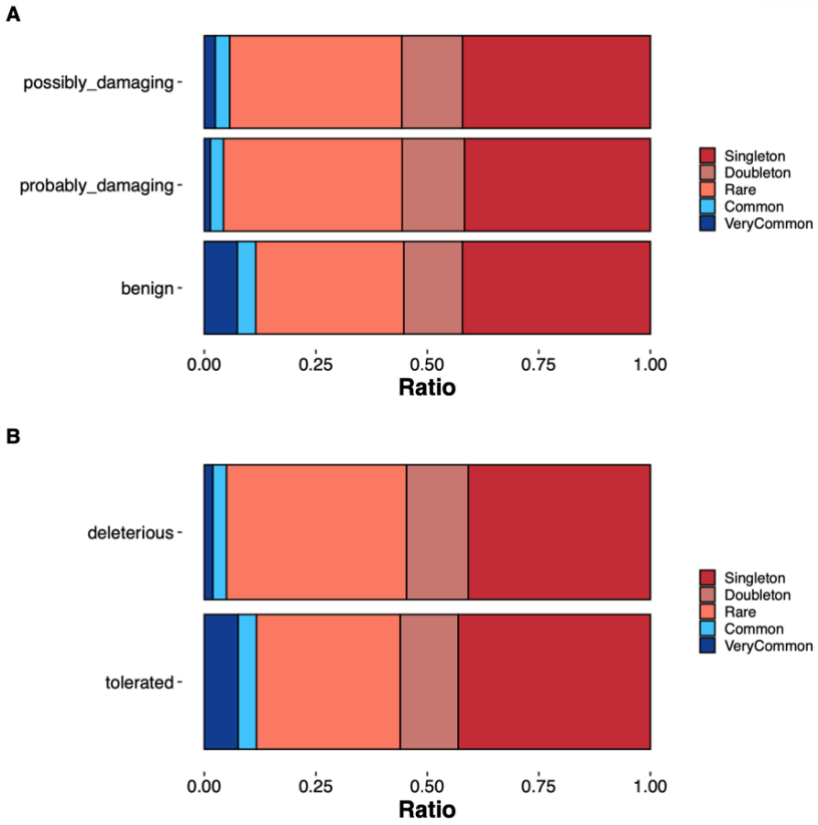


Fig. 12 Proportion of variants based on allele categories for A) PolyPhen and B) SIFT estimation. Singleton: allele count =1; Doubleton: allele count =2; rare: allele count > 2 and allele frequency ≤ 0.01; common: allele frequency > 0.01 and allele frequency ≤ 0.05; very common: allele frequency > 0.05

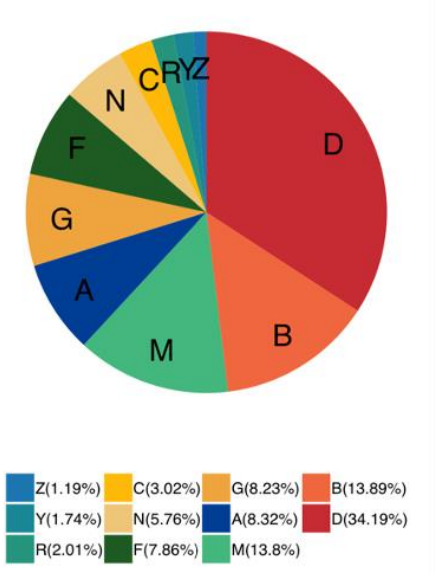


Fig. 13 Mitochondrial haplogroup distribution in Korea1K.

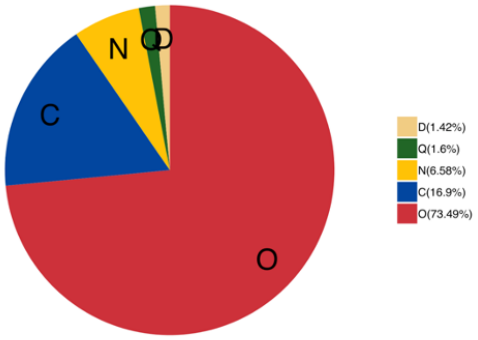


Fig. 14 Chromosome Y haplogroup distribution in Korea1K.

2.3.2. Genomic features of Koreans compared to other populations

I assessed the genetic distinctiveness of Korea1K sample using principal components analysis (PCA) with the small size variants (SNP and indel) in the dataset and 1KGP. As previously reported, principal components PC1 and PC2 with worldwide populations showed a separate East Asian group (Fig. 15A). Although Koreans, Chinese, and Japanese are genetically very close relative to all other individuals[63], I found that those three populations clustered distinctly from each other (Fig. 15B). This pattern was replicated by *ADMIXTURE* analysis with $K = 3$ (Fig 16). To investigate functionally relevant variants, I extracted 1048 ClinVar pathogenic variants found in Korea1K. Among them, 242 variants had an allele frequency greater than 0.1 in Korea1K, which is high for pathogenic variants (Fig. 17). I also found 35 drug-response variants annotated in ClinVar (Fig. 18), and 11 of them displayed significantly different allele frequencies from those of the Chinese or Japanese individuals in the 1KGP set, highlighting the importance of population-specific datasets when interpreting pathogenic or drug-response variants. For example, the variant rs4961 in ADD1 had the highest frequency in the Korea1K compared to other populations and is associated with hypertension and responsiveness to furosemide and spironolactone as shown in a European study[64, 65]. However, no significant association with blood pressure was found in GWAS using the Korea1K set (see the "Genome-wide association study" section for details). I identified CNVs, TE insertions, and HLA-1 haplotypes as WGS data can identify numerous variants from complex or highly variable nongenic regions[26, 27]. In total, 6,131 CNVs were identified in Korea1K (Fig. 19). As expected, since copy number variants are quite variable, more than 50% of the CNVs were categorized as very rare (sample frequency < 0.001). Korea1K contains 1441 CNV loci, and 80% of them overlapped with the CNV set from the entire 1KGP samples while not overlapping with segmental duplication regions. Four common CNVs (sample frequency of >0.05) overlapped with those in the 1KGP set and were validated by a secondary CNV caller, which, in turn, contained five protein-coding genes (Figs. 19 and 20, *LCE3C*, *LCE3B*, *MRGPRX1*, *OR52N5*, and *CLPS*). Among the four common CNVs, Korea1K had a copy duplication of *CLPS*, a pancreatic colipase, which is involved in dietary lipid hydrolysis.

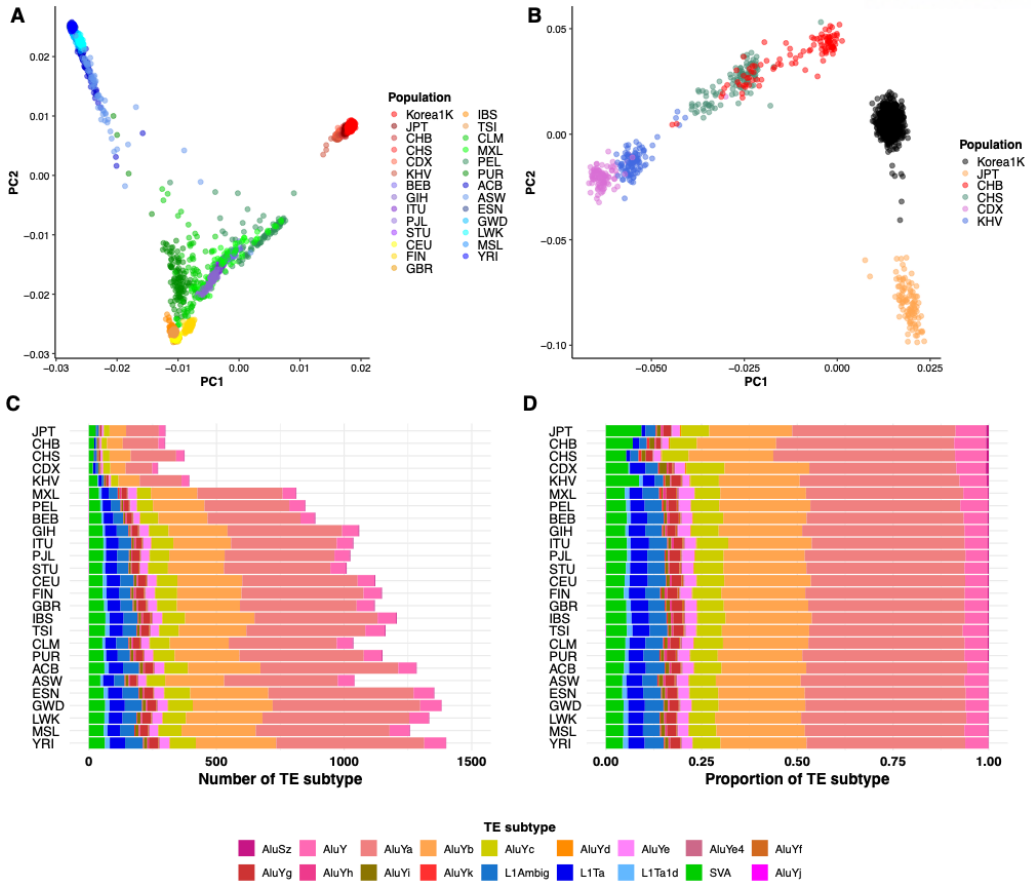


Fig. 15 Comparison with other populations. Results of principal component analysis of Korea1K and the 1KGP set of (A) worldwide populations and (B) East Asian samples. (C) The number of TE insertions with significantly different allele frequencies between the Korea1K set and the population. (D) The proportion of differential TE insertions. Colors indicate TE subtypes. Abbreviation for populations is same population code as 1KGP (ACB: African Caribbean; ASW: African Ancestry SW; BEB: Bengali; CDX: Dai Chinese; CEU: CEPH; CHB: Han Chinese; CHS: Southern Han Chinese; CLM: Colombian; ESN: Esan; FIN: Finnish; GBR: British; GIH: Gujarati; GWD: Gambian Mandinka; IBS: Iberian; ITU: Telugu; JPT: Japanese; KHV: Kinh Vietnamese; LWK: Luhya; MSL: Mende; MXL: Mexican Ancestry; PEL: Peruvian; PJL: Punjabi; PUR: Puerto Rican; STU: Tamil; TSI: Toscani; YRI: Yoruba).

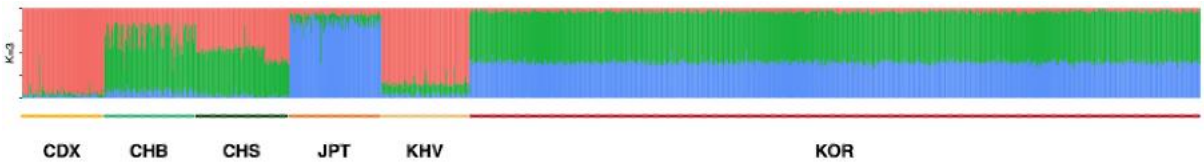


Fig. 16 *ADMIXTURE* plot for Korea1K and 1KGP East Asians. I used K=3 which showed the smallest cross-validation error. (CDX: Dai Chinese; CHB: Han Chinese; CHS: Southern Han Chinese; JPT: Japanese; KHV: Kinh Vietnamese)

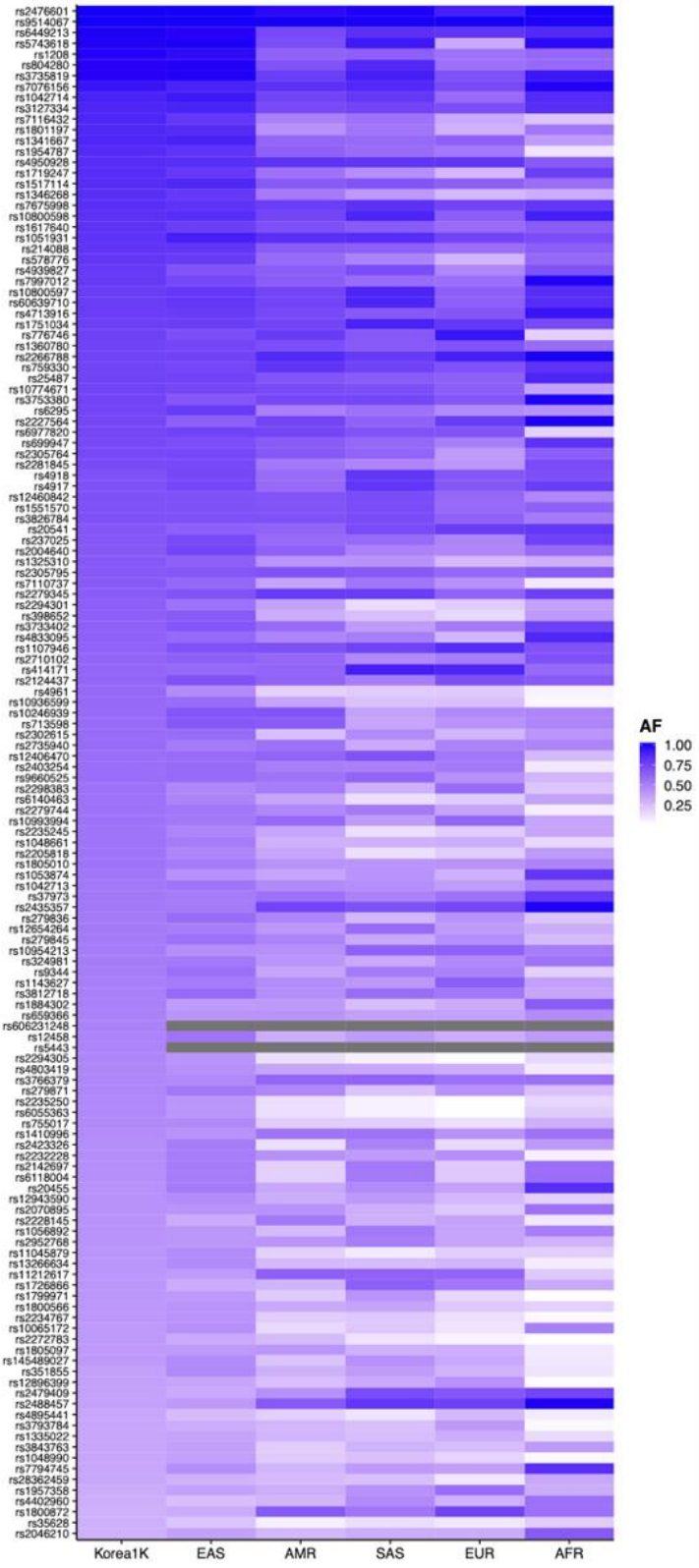


Fig. 17 ClinVar variants which have more than 10% of allele frequency in the Korea1K. The allele frequencies for the super population of 1KGP were also presented. (EAS: East Asian; SAS: South Asian; EUR: European; AMR: American; AFR: African)

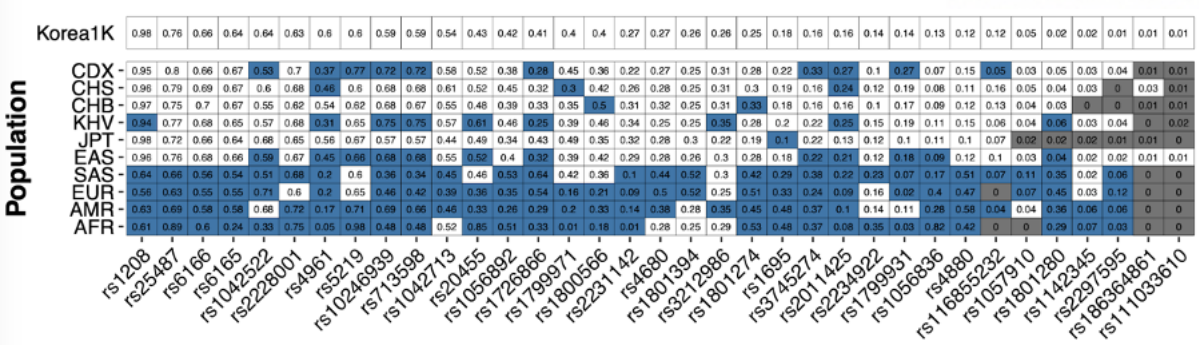


Fig. 18 Drug response variants found in Korea1K. Blue indicates significantly different allele frequencies between the Korea1K dataset and the population from the Chi-square test. White indicates not significant. Grey indicates a Chi-square test could not be performed because of low allele count. Abbreviation on Y-axis is the same population code as 1KGP (CDX: Dai Chinese; CHB: Han Chinese; CHS: Southern Han Chinese; JPT: Japanese; KHV: Kinh Vietnamese; EAS: East Asian; SAS: South Asian; EUR: European; AMR: American; AFR: African).

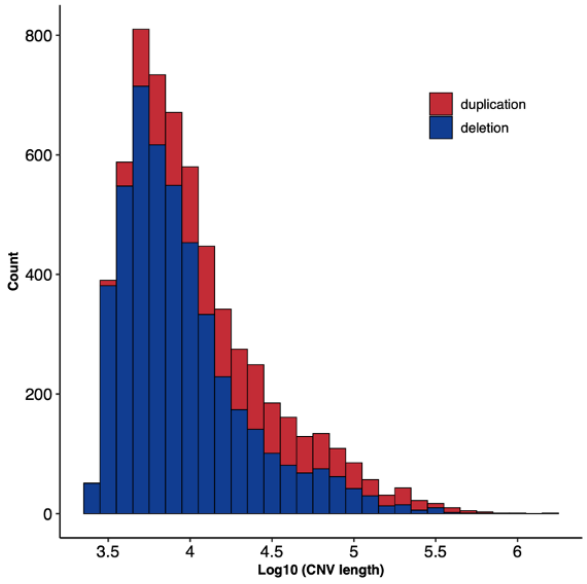


Fig. 19 Length distribution of copy number variations.

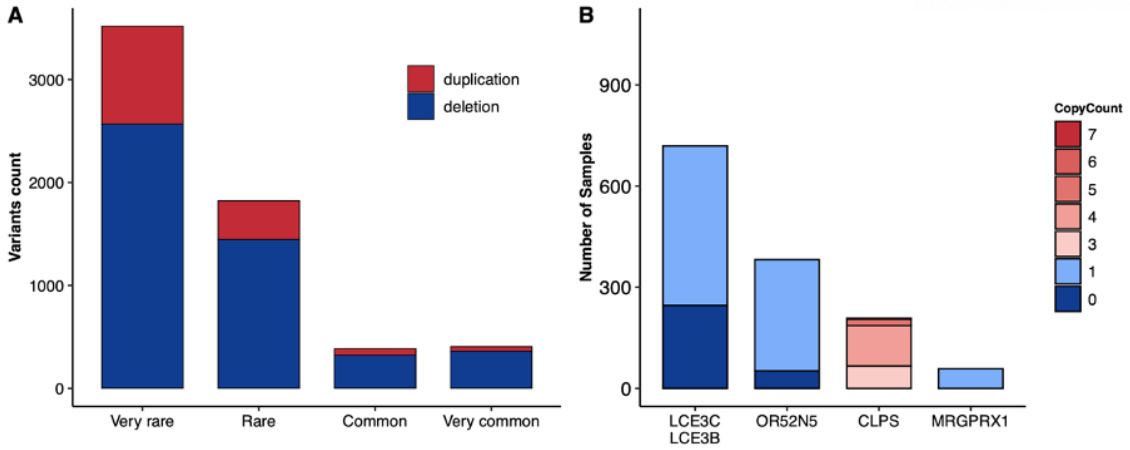


Fig. 20 Copy number variations in Korea1K. (A) The number of copy number variations (CNVs) based on categories of sample frequency. Very rare: sample frequency ≤ 0.001; rare: sample frequency > 0.001 and sample frequency ≤ 0.01; common: sample frequency > 0.01 and sample frequency ≤ 0.05; very common: sample frequency > 0.05. The colors indicate the types of CNVs. (B) Common CNVs overlapped with 1KGP set and protein-coding genes. Colors indicate the copy number.

For TE polymorphisms, in total, 29,143 TE insertions were identified in Korea1K (Alu: 23,915, LINE1: 3,707, SVA: 1,521) from the WGS data (Table 2). More than 50% of the TE insertions identified in Korean1K were rare variants (allele frequency <1%, 16,225 TE insertions, Fig. 21); this pattern was similar to that of SNVs and indels. Allele frequencies of TE insertions were compared between Korea1K and 26 other populations from the 1KGP phase 3 data[38, 41, 66].

Table 2 Number of Transposable element (TE) insertions before and after filtering.

TE type	Number of TE loci before filtering	Number of TE loci after filtering
ALU	23,924	23,915
LINE1	3,708	3,707
SVA	1,522	1,521
Total	29,154	29,143

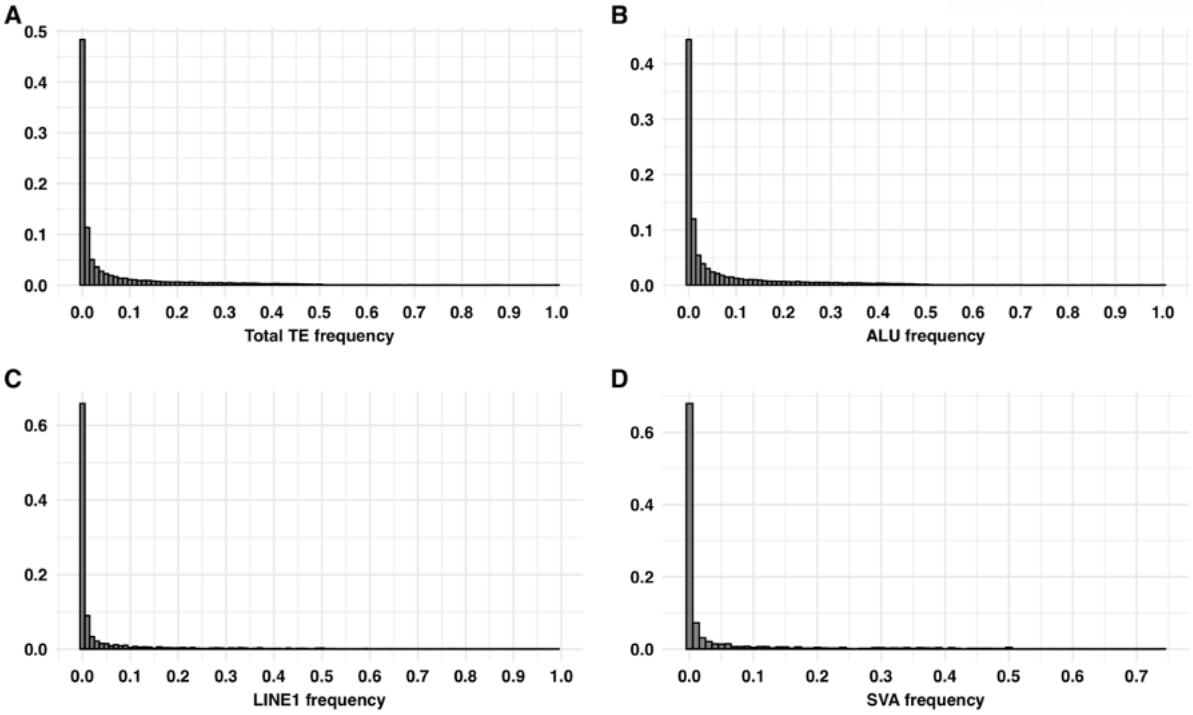


Fig. 21 Transposable element (TE) insertion frequency distribution in Korea1K. (A) All TE type. (B) ALU. (C) LINE1. (D) SVA.

The patterns of TE insertions between the Korean and other populations were investigated by PCA (Figs. 21 and 22 and Table 2). PC1 and PC2 identified that four superpopulations (Africans, Asians, Americans, and Europeans) were well separated from each other, whereas subpopulations in East Asia were not. Therefore, a specific TE insertion pattern alone is insufficient to finely differentiate subpopulations in East Asia, although the genomic diversity was clearly reflected in the allele frequency distribution (Figs. 23 and 24). TE insertions with significantly different allele frequencies between Koreans and 26 other populations in the 1KGP set were enumerated, and as expected, Korea1K displayed significantly fewer differential TE insertions compared to East Asian populations than non–East Asians (Fig. 15C, and D). Furthermore, ALU and SINE-VNTR-ALUs (SVA) displayed a greater proportion of differential TE insertions than long interspersed nuclear element (LINE) in JPT, CHB, and CHS, probably because of different insertion rates on the TE types.

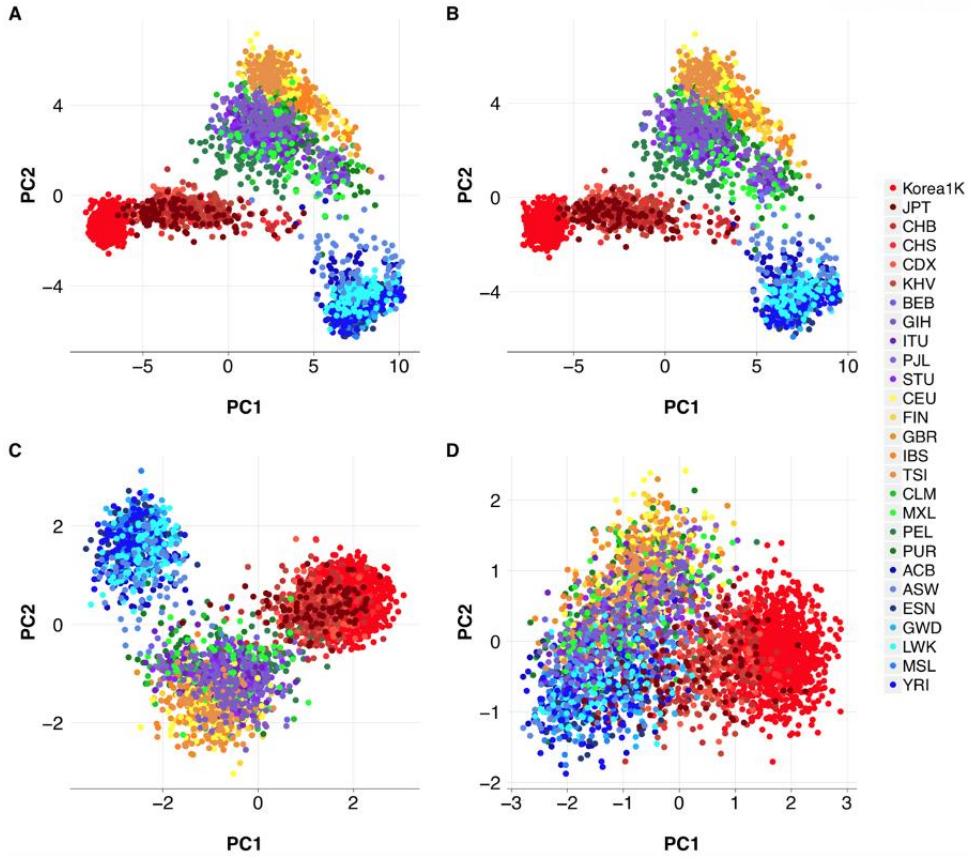


Fig. 22 PCA plot using Transposable element (TE) insertion. (A) All TE types. (B) ALU. (C) LINE1. (D) SVA.

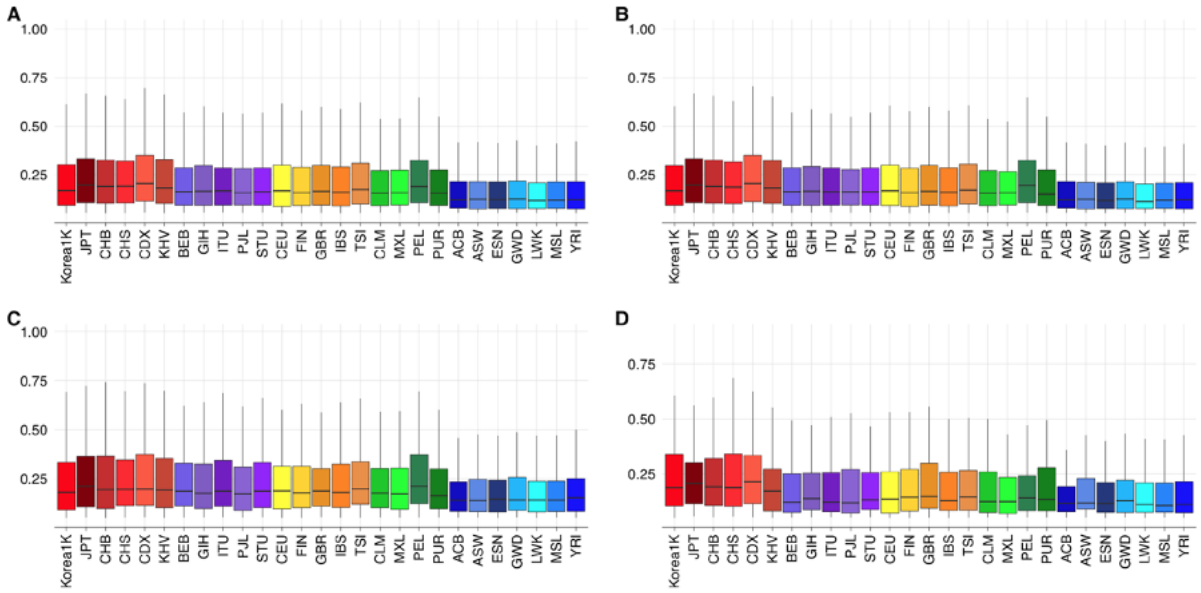


Fig. 23 Transposable element (TE) insertion frequency distribution of Korea1K and 1KGP populations. (A) All TE types. (B) ALU. (C) LINE1. (D) SVA.

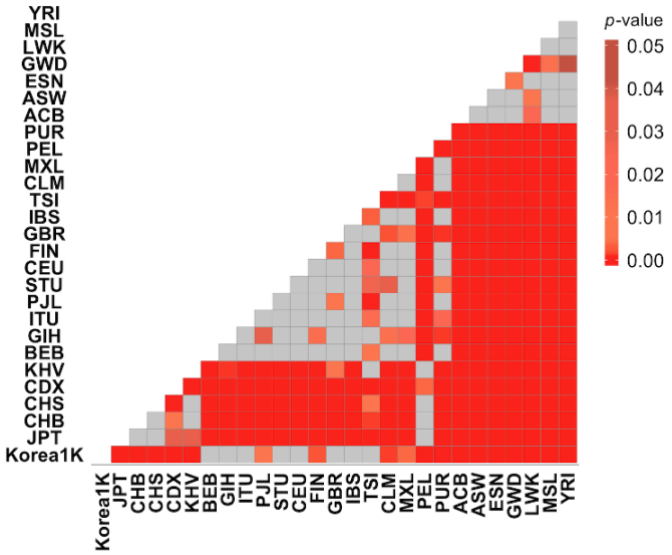


Fig. 24 Significance of TE insertion allele frequency difference. Colors represent P-value. The red box indicates a significant difference in TE insertion allele frequency distribution calculated by the Wilcoxon rank-sum test. The box which does not show a significant difference (P-value >0.05) colored into gray.

I also compared HLA types in Korea1K with those in publicly available databases containing European, American, and Asian HLA frequencies (Figs. 25 and 26). The HLA allele frequency pattern was very similar to the HLA haplotype distribution of Korean samples from the public database. A*24:02, B*44:03, and C*01:02 were the most dominant types of HLA-A, B, and C, respectively (Fig. 25). HLA types A*24:02, A*26:01, A*31:01, B*40:02, and B*52:01 displayed significantly lower allele frequencies in the Korean population relative to the Japanese population (Fisher's exact test $P = 3.61 \times 10^{-49}$, 7.09×10^{-8}, 1.34×10^{-12}, 9.61×10^{-12}, and 3.13×10^{-42}, respectively), while types A*33:03 and B*44:03 had higher allele frequencies (Fisher's exact test $P = 3.10 \times 10^{-46}$ and 1.00×10^{-5}, respectively). Although the Japanese are genetically very close to the Korean, the HLA-type profiles of these populations are considerably different. However, I identified similarities in the Asian populations; for example, types A*33:03, A*02:06, and B*58:01 displayed relatively high allele frequencies, while types A*02:01, A*03:01, A*01:01, A*32:01, A*68:01, B*07:02, B*44:02, and B*08:01 displayed low frequencies in Asian populations (Korean, Japanese, and Chinese populations) compared to other groups.

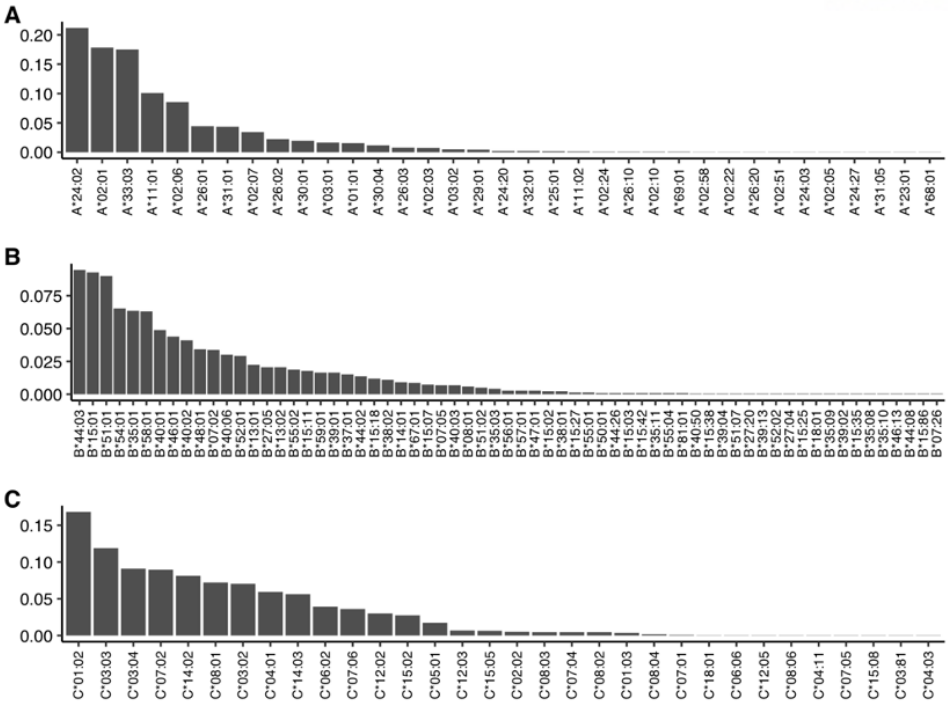


Fig. 25 HLA allele distribution in Korea1K. (A) HLA-A. (B) HLA-B. (C) HLA-C. The X-axis indicates the HLA allele type. Y-axis indicates the proportion of each type in Korea1K.

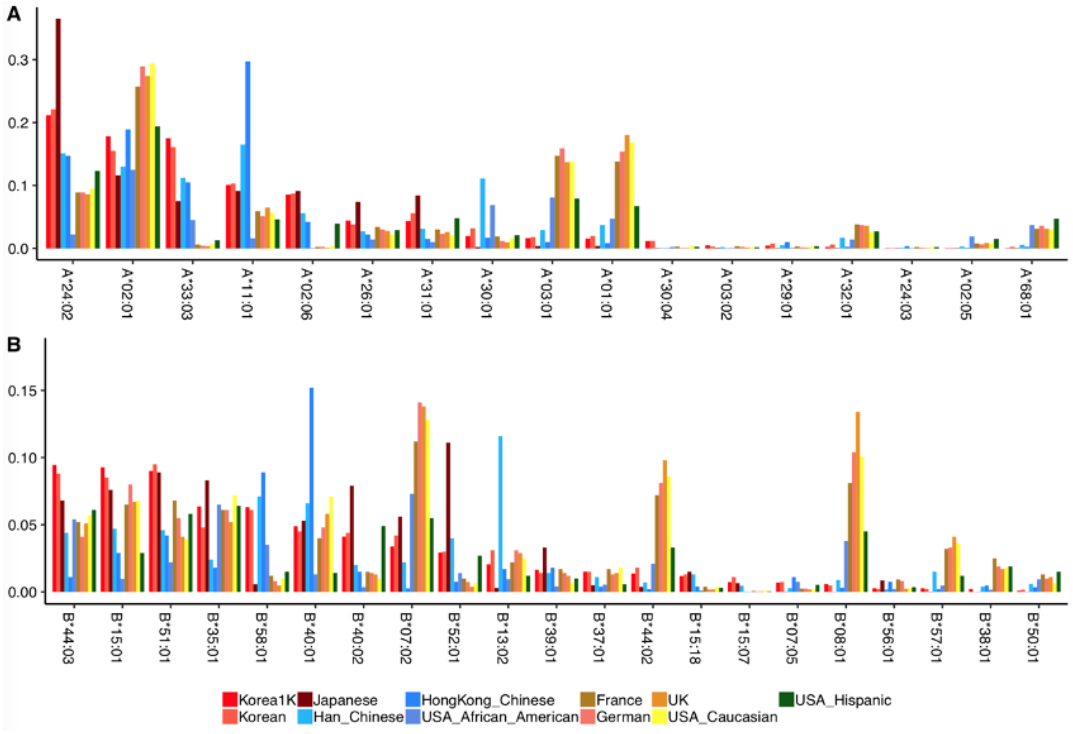


Fig. 26 Comparison of HLA type frequency to the public database. (A) Allele frequency of HLA-A loci. (B) Allele frequency of HLA-B loci.

2.3.3. GWAS based on clinical traits

Thanks to its extensive genomic coverage, population-scale WGS data are more effective than chip-based approaches at identifying statistically significant associations with quantitative traits and diseases[22]. It is even more powerful if matched clinical or phenotypic data are available for the genomes[67]. In the Korea1K set, I was able to quantify 79 quantitative clinical traits measured from 984 samples through health checks provided to the KGP participants. I analyzed associations by fitting additive genetic models with relevant covariates for 79 quantitative traits and 6,658,227 variants [5,932,215 SNVs and 726,012 indels; minor allele frequency (MAF) of >1%] from 823 unrelated individuals of the 984 samples. The analysis resulted in 467 variants that were statistically linked via GWAS to 11 quantitative traits ($P < 7.5 \times 10^{-9}$, the Bonferroni-corrected significance threshold). The 467 variants were clumped into 15 independent loci on eight chromosomes, and 11 of them contained previously reported variants linked to a trait. I found that 11 index variants were not present on the commonly used Illumina Omni 2.5 human SNP chip (Fig. 27-31, Table 3). Among the 11 loci of reported variants, 9 contained variants reported in the GWAS catalog [52], but their index variants were newly identified in this study.

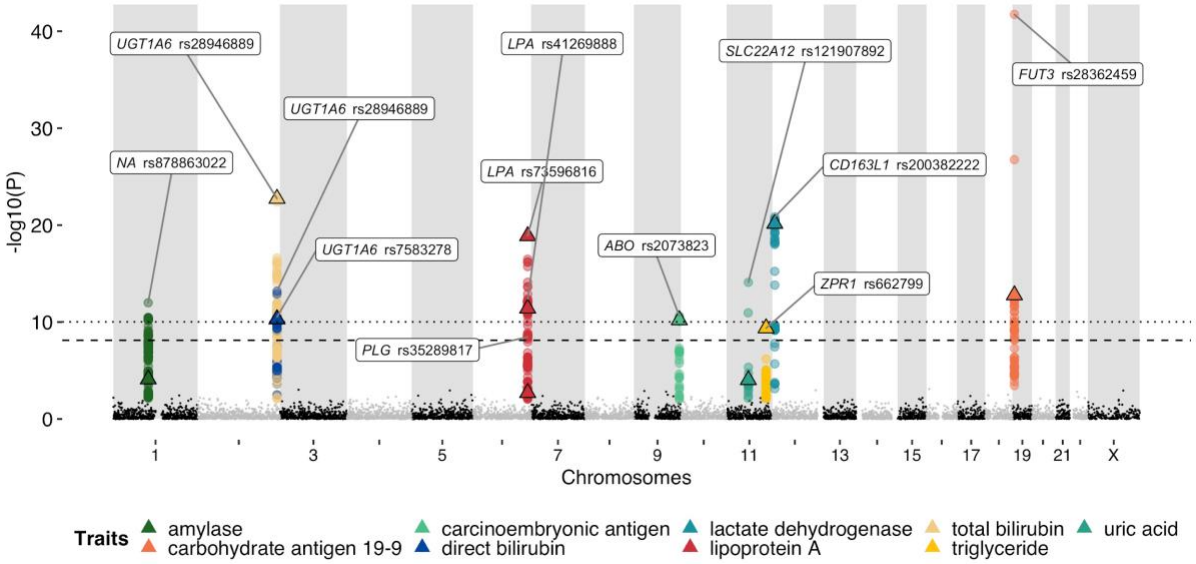


Fig. 27 Manhattan plot of the reported loci via a genome-wide association study. Each color indicates a different clinical trait. The most significant reported markers in the loci are denoted with triangles. The dashed line indicates the threshold for genome-wide significance (7.5×10^{-9}). The dotted line indicates the threshold for study-wide significance (9.5×10^{-11}).

Table 3 List of traits with index variants located in previously reported loci. Highlighted rows indicate unreported variants with higher significance values, located in the same linkage disequilibrium block with reported variants. MAF indicates minor allele frequency.

Trait	Chromosome	Position	rsID	Gene symbol	P-value	MAF
Carbohydrate antigen 19-9	chr19	5,844,781	rs28362459	*FUT3*	1.83E-42	0.341
Total bilirubin	chr2	233,762,816	rs28946889	*UGT1A6*	1.85E-23	0.439
Lactate dehydrogenase	chr12	7,437,350	rs200382222	*CD163L1*	1.40E-21	0.186
Lipoprotein A	chr6	160,596,331	rs73596816	*LPA*	1.31E-19	0.038
Uric acid	chr11	64,593,747	rs121907892	*SLC22A12*	7.94E-15	0.013
Direct bilirubin	chr2	233,762,816	rs28946889	*UGT1A6*	6.43E-14	0.439
Lipoprotein A	chr6	160,607,693	rs41269888	*LPA*	4.30E-13	0.454
Amylase	chr1	103,348,267	rs878863022	*N/A*	1.01E-12	0.476
Carcinoembryonic antigen	chr9	133,257,129	rs2073823	*ABO*	2.53E-11	0.228
Total bilirubin	chr2	233,708,761	rs7583278	*UGT1A6*	2.89E-11	0.100
Neutral fat	chr11	116,792,991	rs662799	*ZPR1*	4.22E-10	0.315
Lipoprotein A	chr6	160,703,093	rs35289817	*PLG*	3.45E-09	0.203

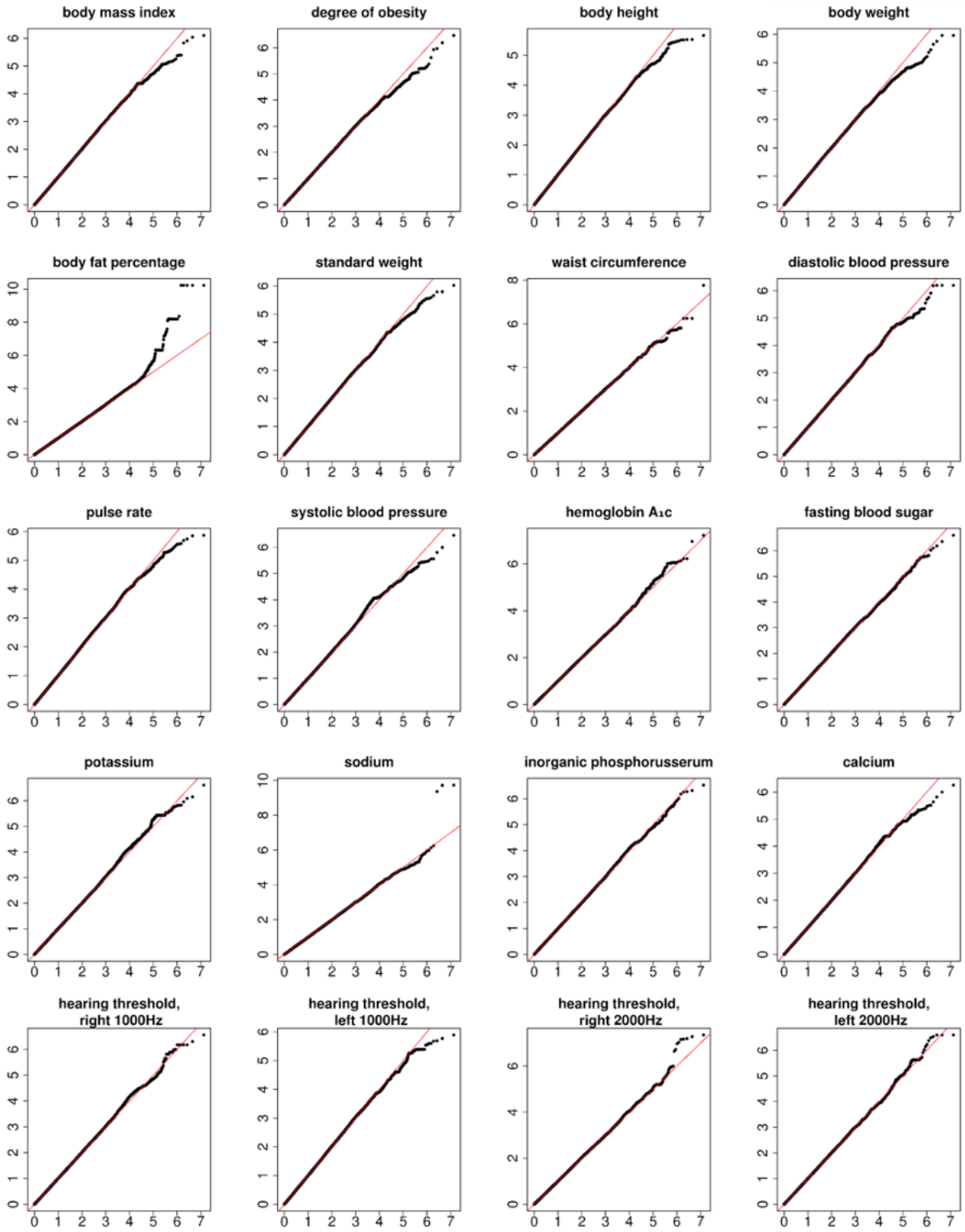


Fig. 28 QQplots for the GWA tests of the 20 traits. X-axis indicates expected -$\log_{10}$ *P*-value. Y-axis indicates observed -$\log_{10}$ *P*-value.

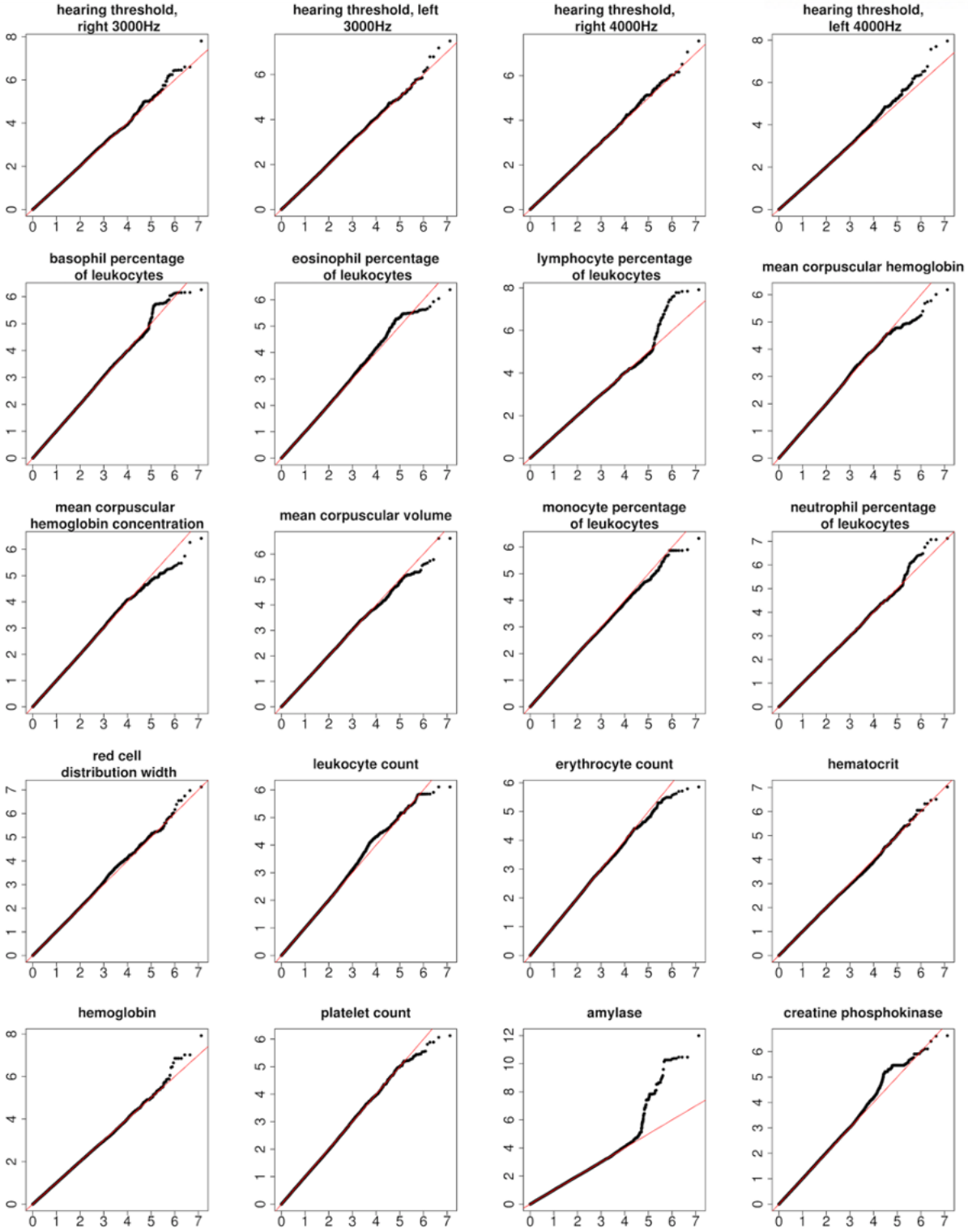


Fig. 29 QQplots for the GWA tests of the 20 traits. X-axis indicates expected -$\log_{10}$ *P*-value. Y-axis indicates observed -$\log_{10}$ *P*-value.

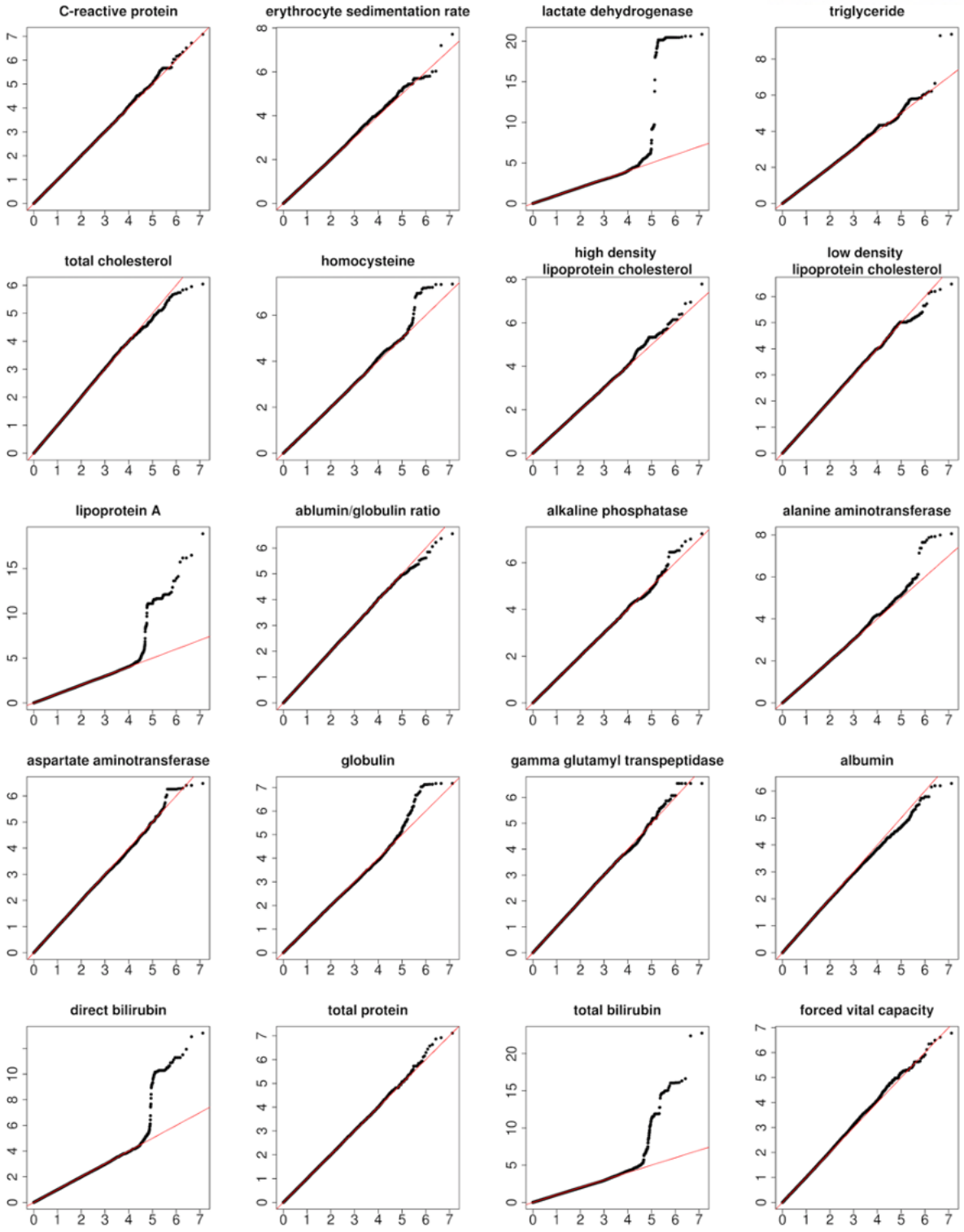


Fig. 30 QQplots for the GWA tests of the 20 traits. X-axis indicates expected -$\log_{10}$ *P*-value. Y-axis indicates observed -$\log_{10}$ *P*-value.

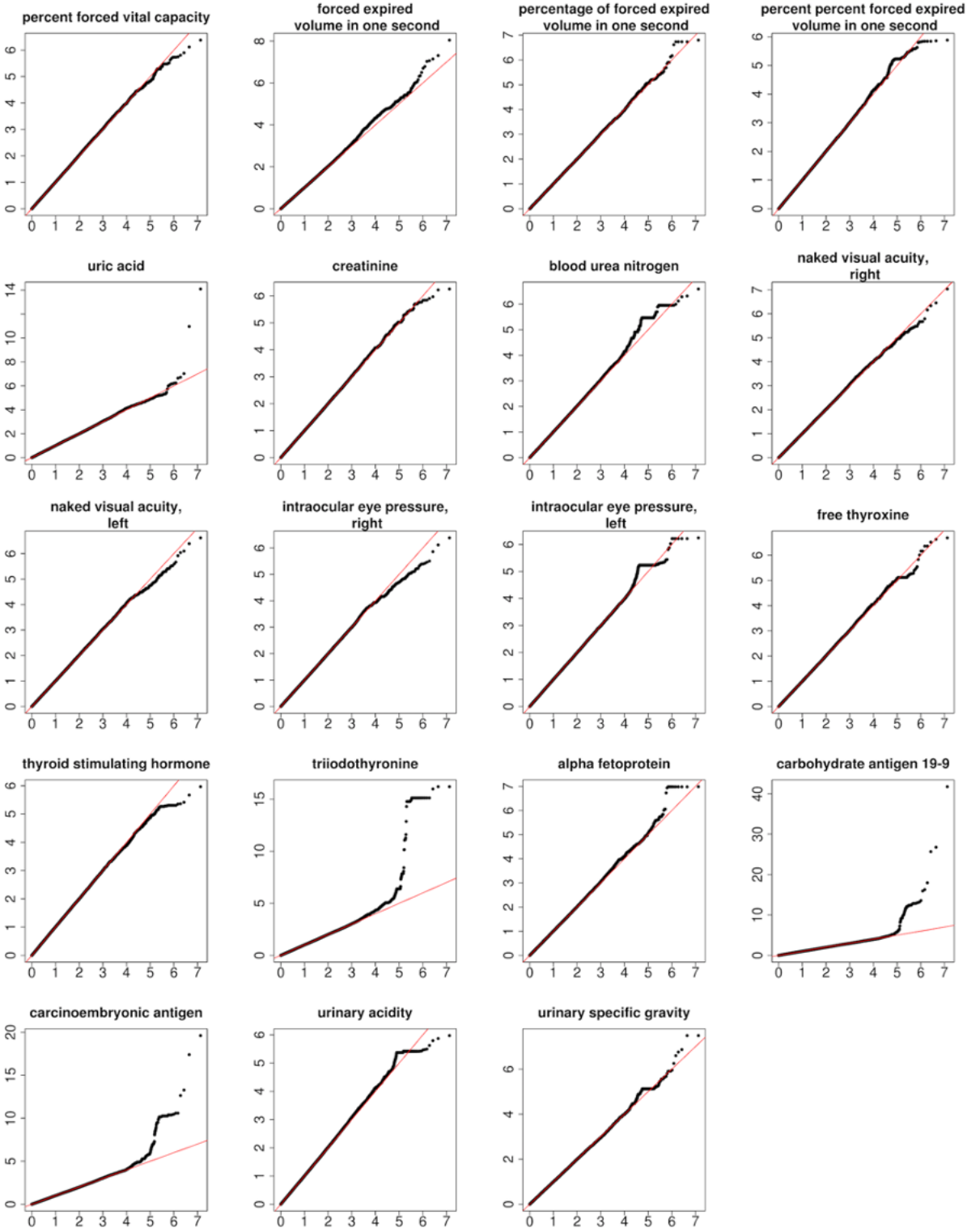


Fig. 31 QQplots for the GWA tests of the 19 traits. X-axis indicates expected -$\log_{10}$ *P*-value. Y-axis indicates observed -$\log_{10}$ *P*-value.

Of the 15 independent loci, I could identify 1 previously unidentified locus in *MAMASTR*, which is associated with cancer markers (carbohydrate antigen 19-9 and carcinoembryonic antigen). A previously unidentified locus was also found in *WDPCP*, which is associated with body fat percentage. Another previously unidentified locus in SERPINA7 was found, which is associated with triiodothyronine. I also found two loci on chromosome 2 (rsID: rs28946889, trait: total bilirubin, $P = 1.85 \times 10^{-23}$; rs662799, neutral fat, $P = 4.22 \times 10^{--10}$) that have been previously identified in two Korean GWAS[68, 69]. The MAFs in the previously unidentified loci were markedly lower than those previously reported when I compared the MAFs of GWAS variants in these previously reported and the unreported loci (Fig. 32). This means that large-scale variomes from WGS data help identify low-frequency alleles and unreported loci via whole genome–based GWA studies.

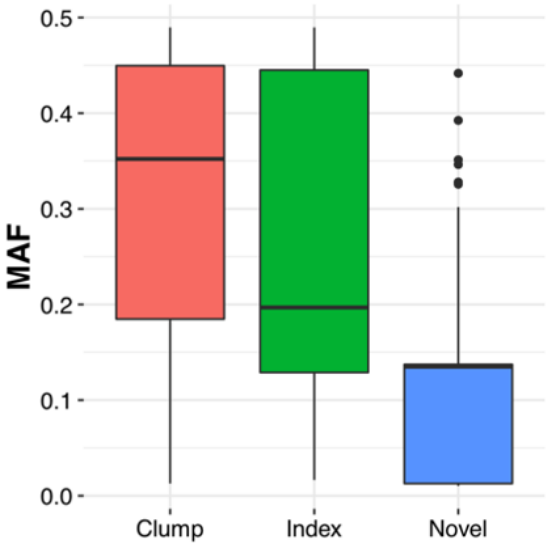


Fig. 32 Minor allele frequency (MAF) of the most significant variant on the loci from GWA analysis. 'Clump' means that clump variants in the loci were reported. 'Index' means that index variants in the loci were reported. 'Novel' means that no variants in the loci were reported.

2.3.4. Korea1K imputation panel

Haplotype-based imputation is a cost-effective method to capture human genetic variation for clinical purposes. Crucially, the accuracy of imputation is improved when a population-specific reference panel is used [27]. I first constructed a phased reference panel using the Korea1K dataset and the combination of Korea1K and 1KGP panel by using SHAPEIT2[53]. The imputation accuracy for the three reference panels [Korea1K (n = 1,059), 1KGP (n = 2,504), and Korea1K + 1KGP (rephased, n = 3,563)] was evaluated by imputing pre-phased variants from the matched normal sample of the 19 Korean patients with gastric cancer. This test set was imputed with the three reference panels using Minimac3[54]. The accuracy was evaluated by comparing the squared Pearson correlation coefficient between the real

genotypes and the dosage of imputed genotypes. The Korea1K panel showed better correlation with true genotypes at low allele frequencies than the 1KGP panel, and the combined Korea1K + 1KGP panel had the best accuracy overall, indicating the usefulness of Korea1K set for the imputation of Korean SNV data (Fig. 33).

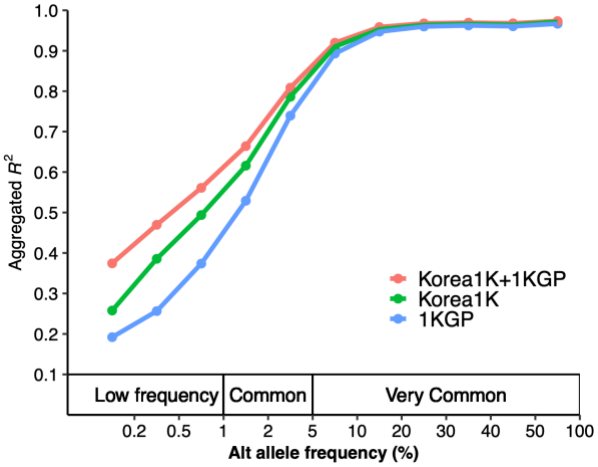


Fig. 33 Imputation performance evaluation. The X-axis indicates alt allele frequency in the Korea1K set. The Y-axis represents the aggregated R^2 values of SNVs. I used SNVs which were overlapped by imputed results across all panels.

2.4. Discussion

In this chapter, I present a comprehensive WGS analysis of 1,094 Koreans (Korea1K), which is a mixture of existing KoVariome[15] and newly added 1,007 genomes with clinical information. On the basis of my analysis, Korean population is genetically homogeneous compared to other East Asians, and this is probably due to geopolitical isolation in the past thousands of years. However, I speculate that, although Koreans are fairly homogeneous, more than 1,000 samples are necessary to map the Korean human genome diversity judging from the assessment of the discovery rate of newly observed variants (allele frequency of >0.05 variants saturated after 132 samples, while the rate of singleton and doubleton variants kept increasing even after analyzing all the 916 healthy unrelated samples). Despite a large amount of genomic data, coupled with clinical information, the CNV and TE analyses did not identify anything unusual or unique. This could be because short-read DNA-sequencing methods have an inherent difficulty in detecting structural variations that cannot be easily resolved bioinformatically, and I must perform long-read sequencing using the same samples in the future to map novel associations between these complex variants and phenotypically accessible traits. Furthermore, I note that the samples are mostly from the Ulsan metropolitan area and cannot reflect the whole Korean peninsula, although Ulsan has a population size of one million and the residents are from all across the peninsula due to rapid industrialization. Together, the sample size of 1,094 from mostly Ulsan is still far from sufficient to represent the Korean population or to map latent genomic structural variations. Nevertheless, the large-scale Korean variome database constructed herein is potentially applicable in studies on various cancers and other diseases of Koreans and can indirectly help reduce the cost of certain genetic analyses. This kind of personal whole-genome dataset combined with common health check–derived clinical information is possibly a good exemplary path for an ethnicity-relevant reference panel for future personalized medical applications for Koreans.

III. Chapter 2. Korean origin

3.1. Introduction

The 1000 Genome Project (1KGP) showed that East Asians displayed a common genetic bottleneck with non-African humans around the last glacial maximum[9]. However, the 1KGP project includes only five EA populations failing to fully represent EA genome structures. In 2009, the HUGO Pan-Asian Consortium (PASNP) confirmed a general concordance between linguistic and genetic affiliations[70]. Most recently, the Asian diversity project showed a correlation between geographical coordinates and genetic structure in Asia[71]. Although Koreans are similar to the Chinese, the PASNP, 1KGP, and Asian diversity projects cannot fully explain the detailed makeup and peopling of the Korean Peninsula.

There are currently several hypotheses on the origins of the Korean. The Korean Y-chromosome haplogroup (O2b-SRY465) suggests the ancestors of the proto-Koreans are related to the people who inhabited northeastern China during the Neolithic (9,900–10,000 years BP) and Bronze (3,450–2,350 years BP) Ages[72]. On the other hand, mitochondrial DNA (mtDNA) shows that Koreans display a very typical East Asian[73]. Previous population studies have revealed that Koreans have not undergone any severe genetic bottlenecks and primarily consist of two genetic components[74]. One is strongly associated with China, but the other is less clear. Therefore, uncovering the exact genetic makeup of Koreans has not been carried out at a whole-genome scale using both present-day and ancient genomes.

Paleogenomics is a powerful tool to reveal the exact genetic lineages and affinities that cannot be resolved with present-day populations alone because frequent and complex genetic exchanges occur with or without cultural and linguistic exchanges. Archeological data unearthed in Korea provide the proto-Korean chronology and prehistories of the Korean Peninsula. The oldest archaic relics, such as the Acheulean axes, that have been found in South Korea date back hundreds of thousands of years, however, human bone preservation is poor due to the acidic soils and cannot acquire any ancient genetic data[75]. The earliest hominid evidences in the Peninsula date to be between 400,000 and 600,000 years ago (YA)[76]. In spite of the claims about human bones in North Korea[75, 77], these paleoanthropological materials are rare in Korea. Therefore, it is only possible to infer the exact Korean ethnic origins through ancient genomes found in the nearby regions, such as Devil's Gate in Russian Far East (8,000 years BP)[11] and Tianyuan cave, Beijing (40,000 years old)[78]. Fortunately, Neolithic to Iron Age ancient genomes from Southeast Asia (SEA) have become available recently[79]. Such ancient genomes, taken from a wide geographic and temporal distribution, should allow us to answer when and how the genomes of Southeast Asia contributed to the genetic makeup of Koreans.

3.2. Methods

3.2.1. Dataset

A total of 88 Korean samples were used that are available from the KoVariome database[15] and 208 worldwide present-day individual samples were collected: 13 African, 4 American, 26 European, 7 Oceanian, 5 Central Asian, 43 East Asian, 31 North Asian, 36 South Asian, 22 West Asian, and 21 Southeast Asian. I collected and added six EA and nine SEA individuals. I merged the whole-genome sequence (WGS) data with the human origin SNP panel data set[80] including six Korean samples' genotype information generated from this panel. A total of 155 ancient genomes were additionally collected. The samples were further chosen to abundantly reflect the target Asian populations and resolve the genetic relationships between Koreans and other populations. All the 88 Korean samples were collected and sequenced according to the guidelines set by the Institutional Review Board (IRB) of the Genome Research Foundation (GRF). Informed consent for study participation was acquired from all participants by the Korean Life Ethics bill, and all experimental protocols were approved by the GRF IRB. I uploaded them on a web site Asian Genome Data for Korean Origin (http://variome.net/Asian_Genome_Data_for_Korean_Origin).

3.2.2. Whole-genome sequencing and genotyping

Genomic DNA was extracted using a QIAamp DNA Blood Mini Kit (Qiagen, CA) and WGS libraries were constructed using TruSeq DNA sample preparation kits (Illumina, CA). Sequencing was performed using Illumina HiSeq sequencers following the manufacturer's instruction. Low-quality reads were removed by NGSQC-toolkit (ver 2.3.3) with "-l 70 and –s 20" options[81]. Filtered reads were aligned to the human reference genome (hg19) using BWA-MEM (ver. 0.7.8)[31]. I further removed PCR duplicates using MarkDuplicates in Picard (ver. 1.9.2, http://broadinstitute.github.io/picard/) and conducted IndelRealigner and BaseRecalibration using GATK (ver. 2.3.9)[82]. I predicted individual single-nucleotide variants using GATK UnifiedGenotyper[82] with "–heterozygosity 0.0010 -dcov 200 -stand_call_conf 30.0 -stand_emit_conf 30.0" options. To confirm artifacts in the variants merging from various resources which can occur during the production process caused by different sequencing platforms, alignment algorithms, and genotype callers, WGS-based variants were merged with the six Koreans' genotypes generated from the human SNP panel data[80]. Finally, I pruned the panel with linkage disequilibrium information using plink with "–indep-pairwise 200 25 0.4" option[46].

3.2.3. Haplotype analysis

Korean haplotypes were analyzed with YFitter[50] for Y-chromosome and haplogrep[83] for mtDNA haplotypes. To analyze the mtDNA haplotypes of the ancient genomes, I downloaded mitochondrial BAM files of ancient genomes via the European Nucleotide Archive with accession ID of PRJEB14817, PRJEB24939, and PRJEB9021 and GenBank with accession ID of KC417443.1 for the Tianyuan

mitochondrion. Consensus sequences of ancient and modern mitochondrial genomes were generated by Samtools with minimal depth 5. Then, multiple sequence alignment of the consensus sequences was performed by MUSCLE. The phylogenetic tree was constructed by MEGA7 with a Gamma distribution model and pairwise deletion for gap treatment. Divergence time between nodes was calibrated by MEGA7 with the four previously suggested calibration points for A (41,504–51,765), B (35,360–44,929), C (29,615–42,453), and D (41,610–52,388)[84]

3.2.4. Genomic clustering

I used CHROMOPAINTER to infer "chromosome chunks" for each individual for fineSTRUCTURE[85] analysis and clustered 88 Koreans and 208 present-day individuals into 64 genetic groups (Fig. 34). The fineSTRUCTURE produced a homogeneous group of 88 Korean individuals (Fig. 35). In total, I reclustered 185 present-day genomes and 6 Korean genomes using CHROMOPAINTER and fineSTRUCTURE[85]. Using these individuals, I implemented *ADMIXTURE* (ver. 1.23)[47] with $K = 2$–14 (Fig. 36). I generated a dendrogram with each of the *ADMIXTURE* result ($K = 2$–14) using the hcluster function in R. I evaluated the consistency of the *ADMIXTURE* and fineSTRUCTURE results by calculating correlation using the "cor.dendlist" function with the "cophenetic" method in the "dendextend" package in R (Fig. 37). It showed the highest correlation when $K = 10$ (corr. = 0.78). I used the admixture result of $K = 10$, which best represents the genetic cluster analyzed by fineSTRUCTURE. I performed a principal component analysis (PCA) analysis conducted with EIGENSOFT (ver. 6.0.1) smartpca[86].

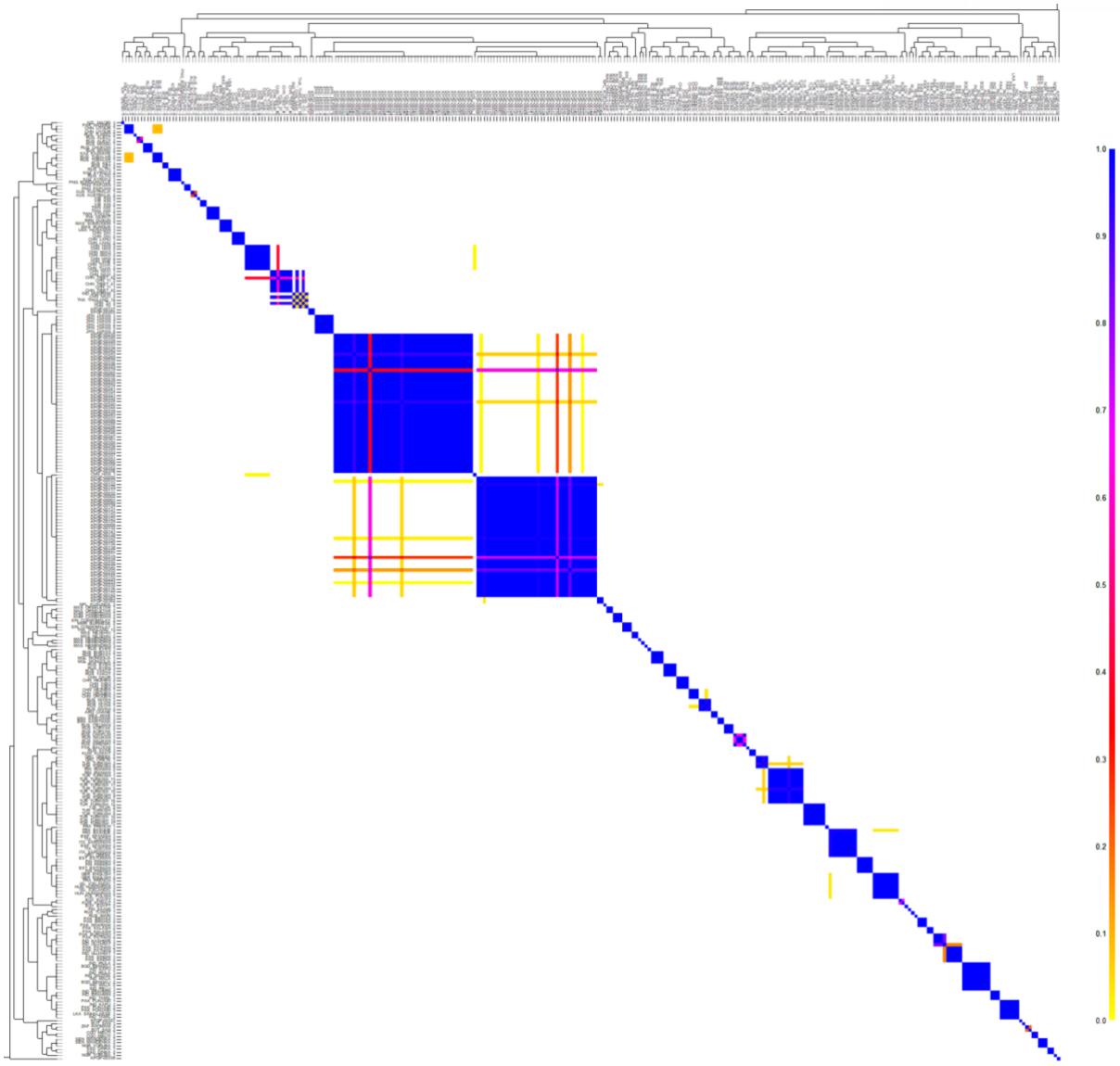


Fig. 34 FineSTRUCTURE analysis of 88 Koreans and 208 contemporary global individuals.

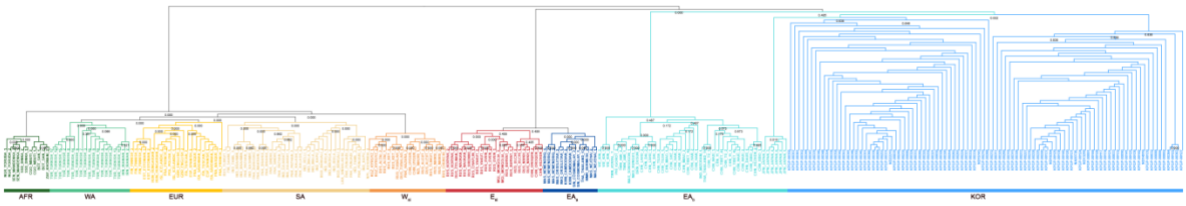


Fig. 35 Global inference of the genetic structures. The phylogenetic tree produced by the fineSTRUCTURE algorithm, which rotates clades according to the geographic associations without topological (branching structure) changes. The genetic distance from the assumed common ancestor is not mathematically scaled and it just shows the cluster topology.

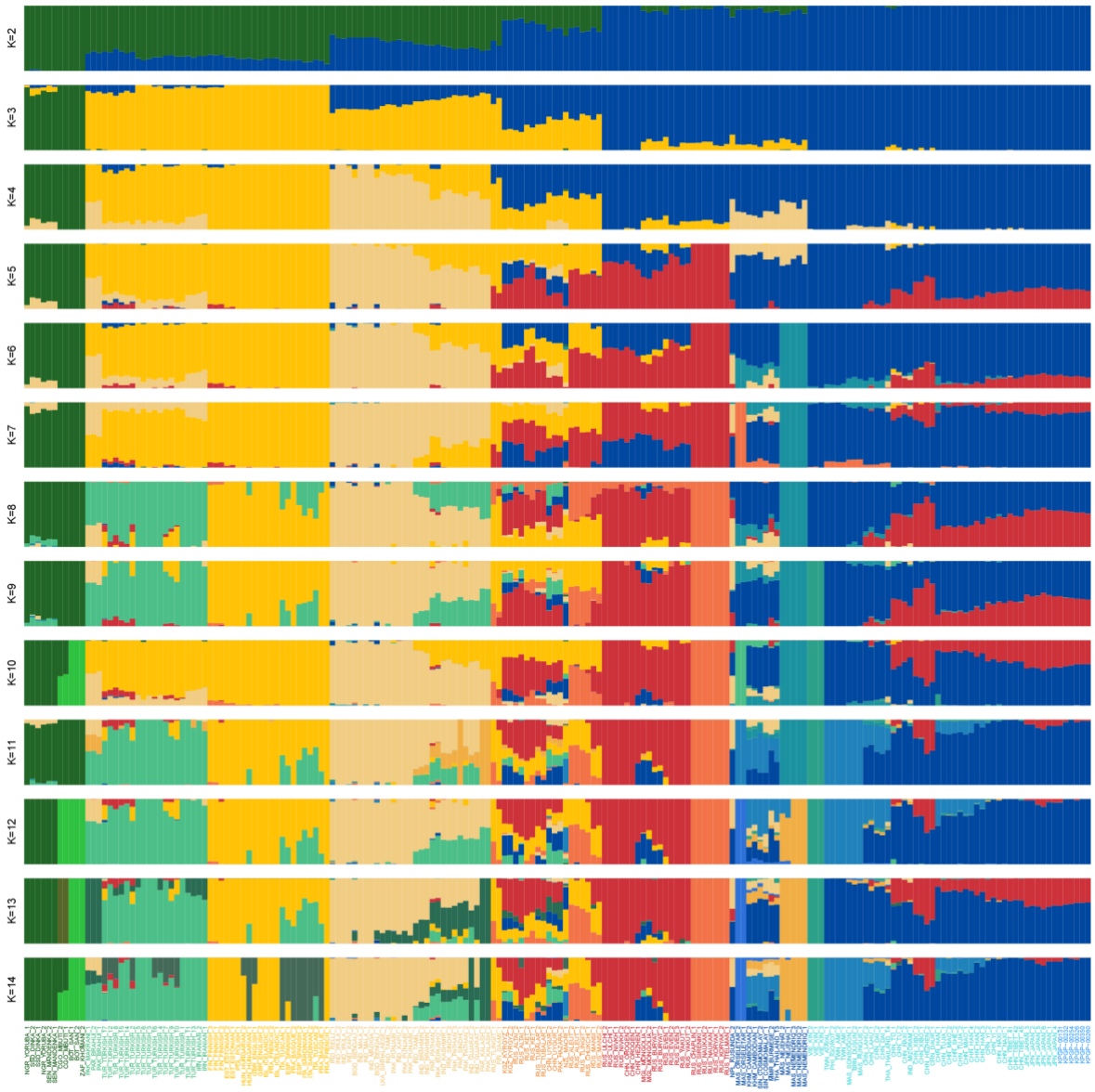


Fig. 36 Global inference of the genetic structures after filtration (*K*=2-14)

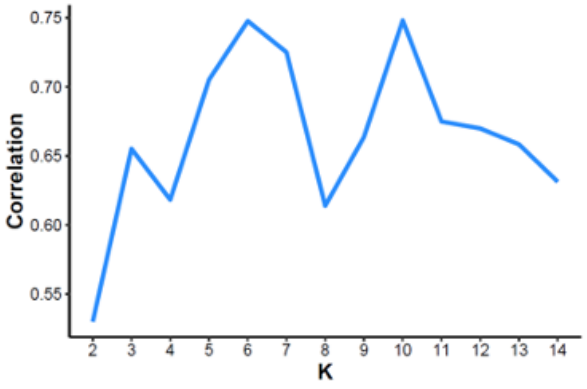


Fig. 37 Dendrogram correlation between the fineSTRUCTURE clade and ADMIXTRUE (*K*=2-14) results.

3.2.5. Admixture time estimation

I implemented the ALDER program[87] to estimate the admixture time of Korean using the Korean itself as one reference population. I used filtering criteria of a genotype rate >99%, MAF > 0.01, and Hardy–Weinberg equilibrium *P*-value > 0.000001.

3.2.6. The genetic affinity between the ancient and present-day populations

To investigate the genetic relationship between populations of interest, I used the *D* and outgroup *f3* statistic framework by using ADMIXTOOLS[88]. The genetic affinity between the ancient and present-day populations was measured with the outgroup *f3* statistic using the following notation: *f3*(X, Y; Yoruba), where X and Y are ancient and present-day populations, respectively. To better represent the genetic association of the present-day population against a focal ancient genome, I applied a scaled *f3* statistic by $f3_{scaled} = (f3 - m)/(M - m)$, where m and M represent the minimum and maximum *f3* statistic. To cluster ancient genomes in this study, I analyzed a pairwise outgroup *f3* statistic with a form of *f3*(X, Y; Yoruba). In this analysis, both X and Y were ancient genomes.

3.2.7. Admixture model construction

To construct an admixture model depicting the historical genetic makeup of Koreans and other Asians, I fitted the SNP panel to the admixture models with the qpgraph program[88] based on results from *D*-statistics and *f3* statistics in the study. I first set the skeleton for the admixture model as Tianyuan, Onge, and Ami by adapting a previous study[89] (worst-fitting Z = 0.044). Then, I added Kinh which has a high admixture *F3* score with Devil's Gate to Koreans (worst-fitting Z=−3.887) and then to Devil's Gate, Ulchi, Koryak, Mixe, and MA1 (worst-fitting Z = 3.317). Finally, Koreans, Han, and Japanese have been added to model the suggested admixture of East Siberians (E_{si}) and East Asians b (EA_b) (worst-fitting Z value of −3.686). I manually calibrated the final model with a time point which was estimated using the ALDER results.

3.3. Results and discussion

3.3.1. Korean genetic structure

To infer the genetic association between the 88 Koreans and the selected neighboring populations, I collected with WGS from 185 contemporary individuals belonging to 91 populations (Fig. 38A).

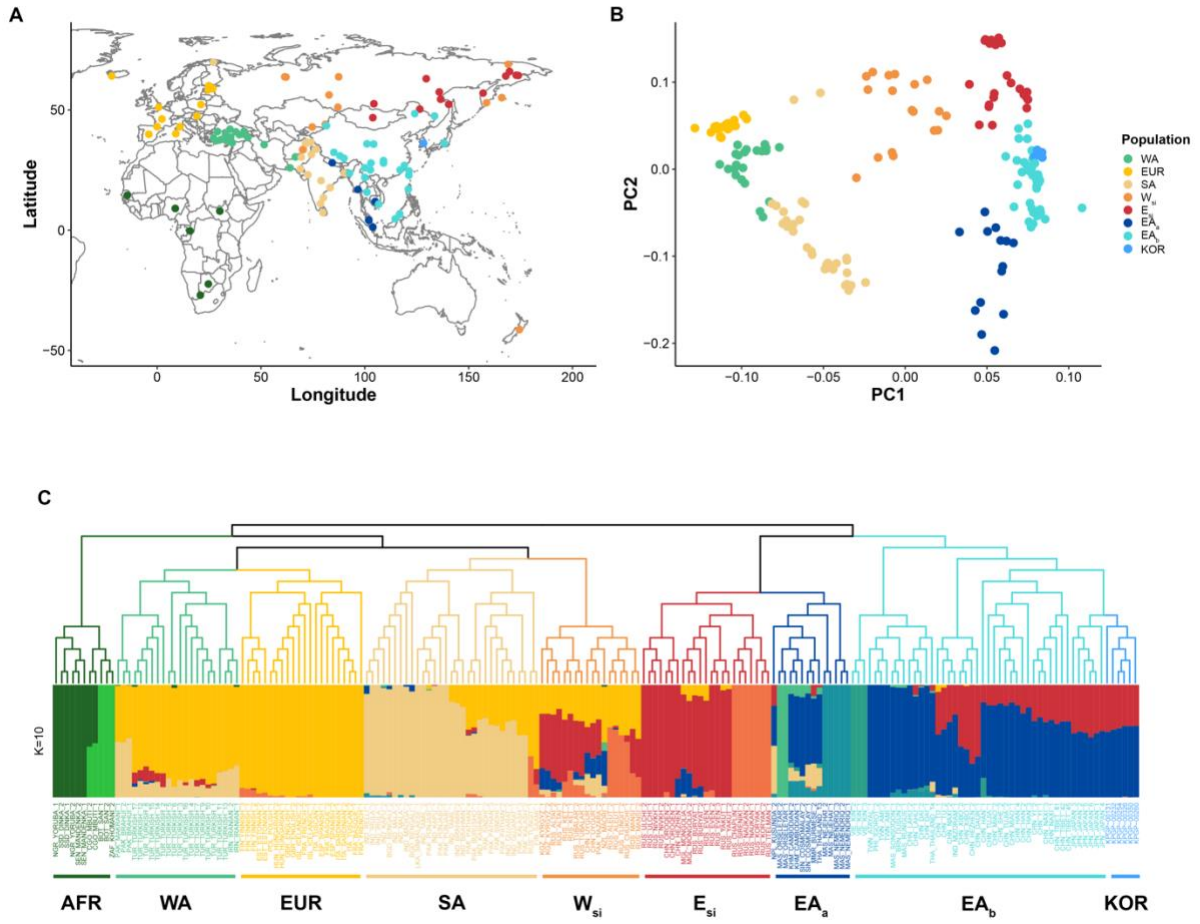


Fig. 38 Genetic clustering of the present-day populations

(A) Illustration of the geographical distribution of the 91 populations analyzed in this study. (B) Principal component analysis (PCA) of the 185 individuals using 199,629 linkage disequilibrium (LD) pruned SNPs in the 109 present-day populations. (C) Genetic clustering of present-day populations analyzed by fineSTRUCTURE [85] (top) and *ADMIXTURE* [47] (bottom). Names of the genetic clusters are given underneath the admixture group names.

I included people from 21 and 31 Southeast Asian and North Asian ethnic groups, respectively, from which Koreans could have originated. I predicted an average of 1.5 and 2.6 mega homo- and

heterozygous single-nucleotide variants from each individual, respectively. I merged WGS-based SNPs with the human origin SNP panel data set and finally produced 199,629 autosomal SNPs for genetic comparison. To infer the genetic structures of the Korean ethnic group, I clustered 94 Koreans, including 6 published Koreans genotyped with SNP chip, by applying the CHROMOPAINT and fineSTRUCTURE[85] programs. These algorithms clustered 279 individuals into 64 homogeneous groups according to the haplotype patterns shared by the individuals (Fig. 34). This analysis showed eight global haplotype patterns: Africans (AFR), West Asians (WA), Europeans (EUR), South Asians (SA), West Siberians (W_{si}), East Siberians (E_{si}), and two groups of East Asians (EA_a and EA_b) (Fig. 35), which reflect both geographic and genetic relationships (Fig. 38A). The group of EA_b consists mainly of Korean, Chinese, Japanese as well as Austroasiatic speakers in Southeast Asia and EA_a contains several ethnic minorities of Southeast Asia. I first confirmed a genetically homogeneous ethnic group of Koreans by showing a single clade in the fineSTRUCTURE tree (Fig. 36). This homogeneity is also consistent across chip-based and WGS-based data, suggesting that there is no technical bias in the sequencing platform or the SNP prediction algorithm. In the PCA, both the Koreans and EA_b fell between the EA_a and E_{si} populations (Fig. 38B), consistent with other previous studies[63, 90]. I reanalyzed fineSTRUCTURE and *ADMIXTURE*[47] with 6 randomly sampled Koreans and 185 global populations, to compare Korean's genetic components without sampling bias (Fig. 38C). Consistent with the PCA result, the fineSTRUCTURE tree showed Koreans formed a homogeneous clade with most of the EA populations represented by EA_b and their sister groups were composed of E_{si} and EA_a (Fig. 38C top). I also analyzed genetic ancestry assuming ancestral groups from $K = 2$ to $K = 14$ in the *ADMIXTURE* analysis[47] (Fig. 36). From $K = 5$, it showed two genetic components, red and blue, were admixed in Koreans which were dominated in the E_{si} and $EA_{a/b}$ populations, respectively; although, these ratios were slightly different depending on the number of ancestral groups (K). The dendrogram correlation analysis showed the greatest consensus between the fineSTRUCTURE clades and *ADMIXTURE* results at $K = 10$ (Fig. 37). At $K = 10$, I observed 38% and 62% of the E_{si} and $EA_{a/b}$ genetic components in the Koreans, respectively (Fig. 38C). Korean and Japanese populations showed very similar levels of genetic admixture rates, consistent with their sister groups in the fineSTRUCTURE tree (Fig. 38C). Takeuchi et al. (2017) reported a high degree of genetic similarity between the Korean and mainland Japanese and the estimated admixture date of the EA-wide genetic component to Japan was in the Yayoi period (3,000–1,700 years BP)[74]. The Chinese also have similar genetic compositions to the Korean and Japanese; however, their admixture rates differed depending on geographic region. Overall, I conclude that genetic admixture events occurred first between the Southeast Asians and Chinese outside Korea and Japan and then spread, rather than occurring separately in Korea or Japan locally. It is also possible that such a recent genetic admixture was a broad phenomenon, happening concurrently all across EA driven by a population expansion caused by the agricultural, economic, and technological advances of the last 4,000 years[79].

3.3.2. The gene flow from Neolithic age Devil's gate ancestry to Korean people

To reveal past genetic exchanges contributing to the current Koreans and their neighboring populations, I collected 115 ancient genomes from across the world, consisting of 4 Pleistocene hunter-gatherers, 13 Holocene hunter-gatherers, 20 Early Neolithic, 10 Mid Neolithic, 10 Late Copper Age, 9 Late Neolithic, 20 Early Bronze Age, 4 Mid Bronze Age, 2 Late Bronze Age, and 12 Iron Age ancient genomes distributed across European and Russian regions. The time scale of these ancient genomes was categorized by referring to previous research[91]. In addition, I included the Tianyuan genome from northern China[78], two ancient genomes unearthed from the Devil's Gate cave near North Korea [11], and eight ancient genomes from Southeast Asia dating from the Neolithic to the Iron Age [79], making a total of 115 genomes. I measured levels of pairwise genetic affinity among the ancient and present-day genomes by using outgroup *f3*-statistics, with a form of *f3*(ancient, present-day; Yoruba)[88]. This analysis calculates the global landscape of the genetic associations between ancient and present-day genomes (Fig. 39). The $f3_{scaled}$-statistics showed that the ancient Tianyuan individual (40,000 years BP from China) shares more alleles with present-day Siberians (E_{si} and W_{si}) and East Asian (EA_b) populations than with other present-day populations such as European, West-, and South Asians (Fig. 39). It suggests Tianyuan is the basal genetic component of the East Eurasian and East Asian lineage. I also observed that present-day E_{si} and EA_b populations had significant genetic affinities with ancient Southeast Asians (ancSEA), Devil's Gate, and Bronze and Iron age ancients who lived in central steppe regions (ancCS) (Fig. 40A).

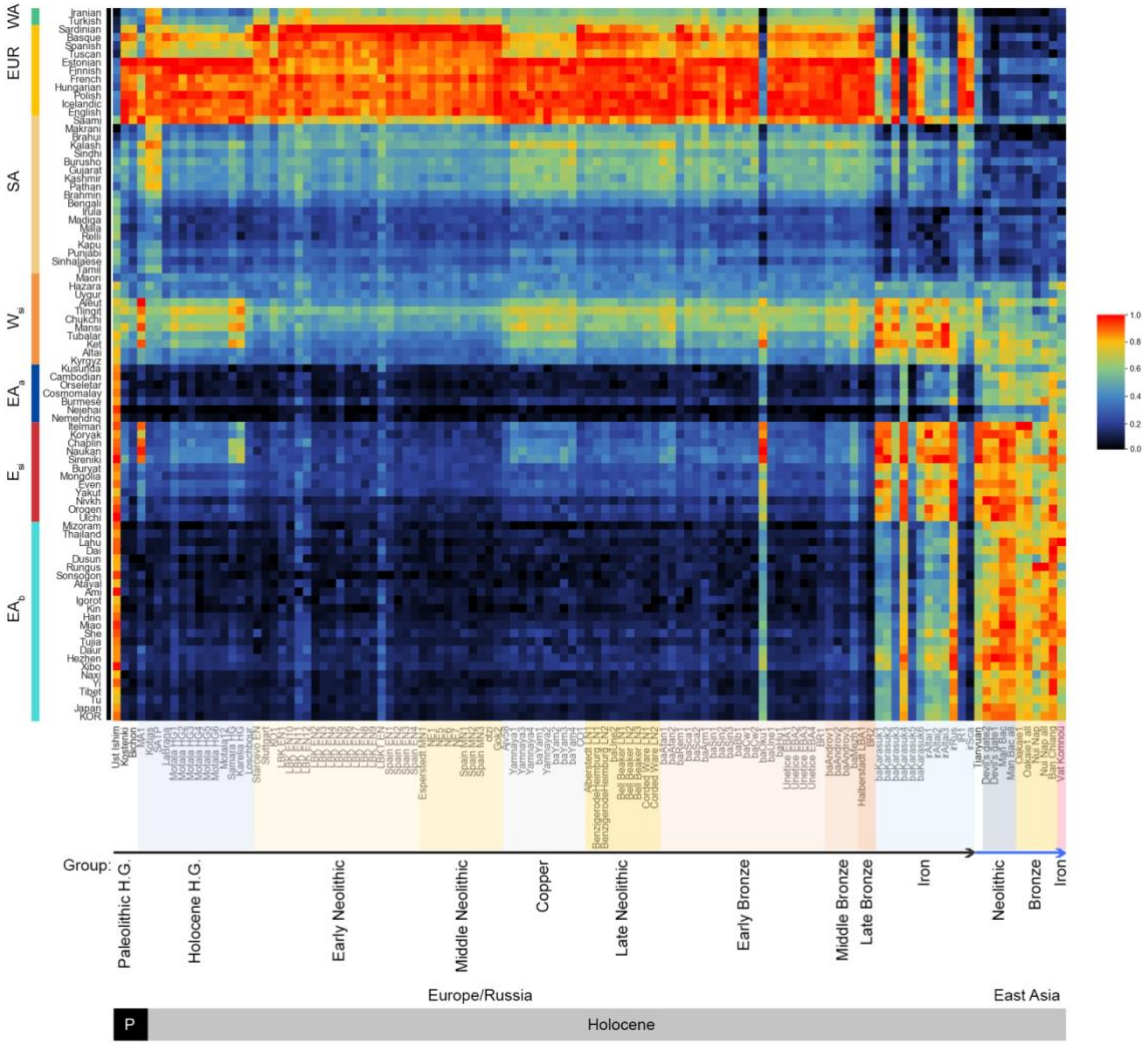


Fig. 39 A global outgroup *f3* statistics between the ancient and present-day populations

Outgroup *f3* analyses with the form of *f3*(X, Y; Yoruba), where X and Y are ancient and present-day populations, respectively. I scaled the *f3* statistics between 0 (black) and 1 (red) in the heat map. For ancient genome X (on rows), the scaled *f3* statistic for a given cell in that column is calculated by $f3_{scaled}$ = (f3 − m)/(M − m), where m and M represent the minimum and maximum *f3* statistic. I ordered ancient genomes in the X-axis according to the time scale.

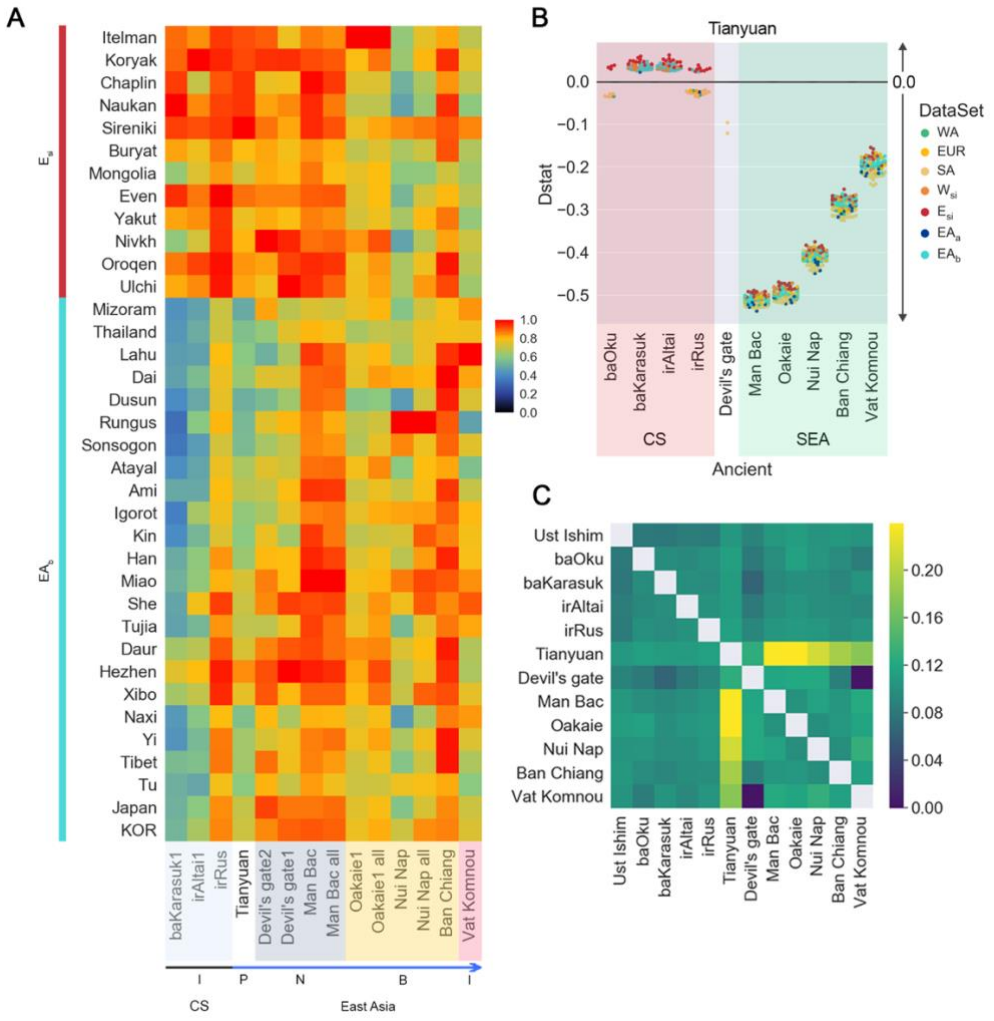


Fig. 40 Genetic association between the ancient and present-day populations.

(A) outgroup *f3* statistics with the form of *f3*(X, Y; Yoruba), where X and Y are ancient and present-day populations, respectively. I scaled *f3* statistics between 0 and 1. In the heat map, black indicates that the $f3_{scaled}$ value is close to 0 and red indicates the value is close to 1. For ancient genome X (on rows), the scaled *f3* statistic for a given cell is calculated by $f3_{scaled} = (f3 - m)/(M - m)$, where m and M represent the minimum and maximum f3 statistic. Therefore, the smallest *f3* in each column has $f3_{scaled}$ statistic = 0 (black) and the largest has f3scaled statistic = 1 (red). I ordered ancient genomes in the X-axis according to the time scale. I also separated Central Steppe (CS) ancestry (black arrow) [92] and Chinese and Southeast Asian ancestry genomes (blue arrow) [79]. Abbreviations: P on the bottom bar represents Pleistocene hunter-gatherers. The N, B, I represent Neolithic hunter-gatherer, Bronze and Iron age, respectively. (B) *D*(Yoruba, Tianyuan; X, Y), where X and Y are ancient and present-day populations, respectively. I represented only the absolute |Z-score| >3. The X-axis represents ancient genomes that have a genetic affinity with East Asia (EA) and East Siberia (E_{si}) populations. (C) outgroup *f3* statistics among ancient genomes with the form of *f3*(X, Y; Yoruba). Both X and Y were ancient genomes.

Based on these genetic affinities, I deduced the genetic founders of the Koreans by comparing the Tianyuan-derived alleles shared with these ancients and present-day populations. I applied *D*-statistics in the form of *D*(Yoruba, Tianyuan; X, Y), where X and Y were ancient and present-day populations, respectively (Fig. 40B and 41). Tianyuan shares more derived alleles with ancSEAs than with any present-day populations (Fig. 40B), suggesting ancSEAs directly come from the Tianyuan lineage. Neolithic Devil's gate and present-day population (E_{si} and $EA_{a/b}$) showed a similar amount of Tianyuan's genetic ancestry by showing *D*(Yoruba, Tianyuan; Devil's Gate, E_{si} or $EA_{a/b}$) ≈ 0. It suggests Neolithic Devil's gate (Northern part of Korea) is possible to be admixed with another genetic component. In addition, Tianyuan's genetic ancestry had a significantly higher level of genetic affinity with W_{si}, E_{si}, and EA_b populations than with ancCS (Fig. 40B). It suggests ancCS were possibly generated from other genetic compounds. The genetic clustering of ancient genomes also confirmed the highest genetic affinity of Tianyuan in Man Bac and a slight reduction of this affinity in other ancSEAs over time (Fig. 40C and 42). This evidence suggests ancSEA received an additional genetic component over time, consistent with Man Bac having the highest affinity toward Tianyuan. I examined Tianyuan's genetic affinities for E*si* and $EA_{a/b}$ using *D*-statistic in the form of *D*(Yoruba, Tianyuan; E_{si}, $EA_{a/b}$) (Fig. 43). In these statistics, the Tianyuan genome showed a higher level of genetic affinity with present-day Esi than Southeast Asians. However, several EA_b (Korean, Japanese, and south Chinese) populations showed similar levels of affinity with Tianyuan-derived alleles to the E_{si} populations and were equally distant to Tianyuan lineage. This suggests Devil's Gate ancients and present-day E_{si} and several EA_b populations were subject to similar genetic influences over time and are expected to be a single clade since they are all separated originally from the Tianyuan lineage. These lines of analysis reveal that the basal ancient of the Tianyuan genome was separated in the Neolithic or pre-Neolithic era and independently affected current Koreans.

Fig. 41 D-statistics with a form of *D*(Yoruba, Tianyuan, ancient, present-day)

D(Yoruba, Tianyuan, ancient, present-day) test suggests Tianyuan is a putative founder of the EA population by showing significantly positive *D*-stats for the E_{si} population but not for the EA_b (none or equal gene flow level). The Tianyuan genome had significantly higher allele sharing with Neolithic than Ion Age Southeast Asian ancient populations compared to present-day EA or Siberian populations and these sharing statistics value decreased over time. It supports the scenario of continuous Tianyuan-derived genetic association with Southeast Asians before and during the Neolithic Age until the Bronze Age.

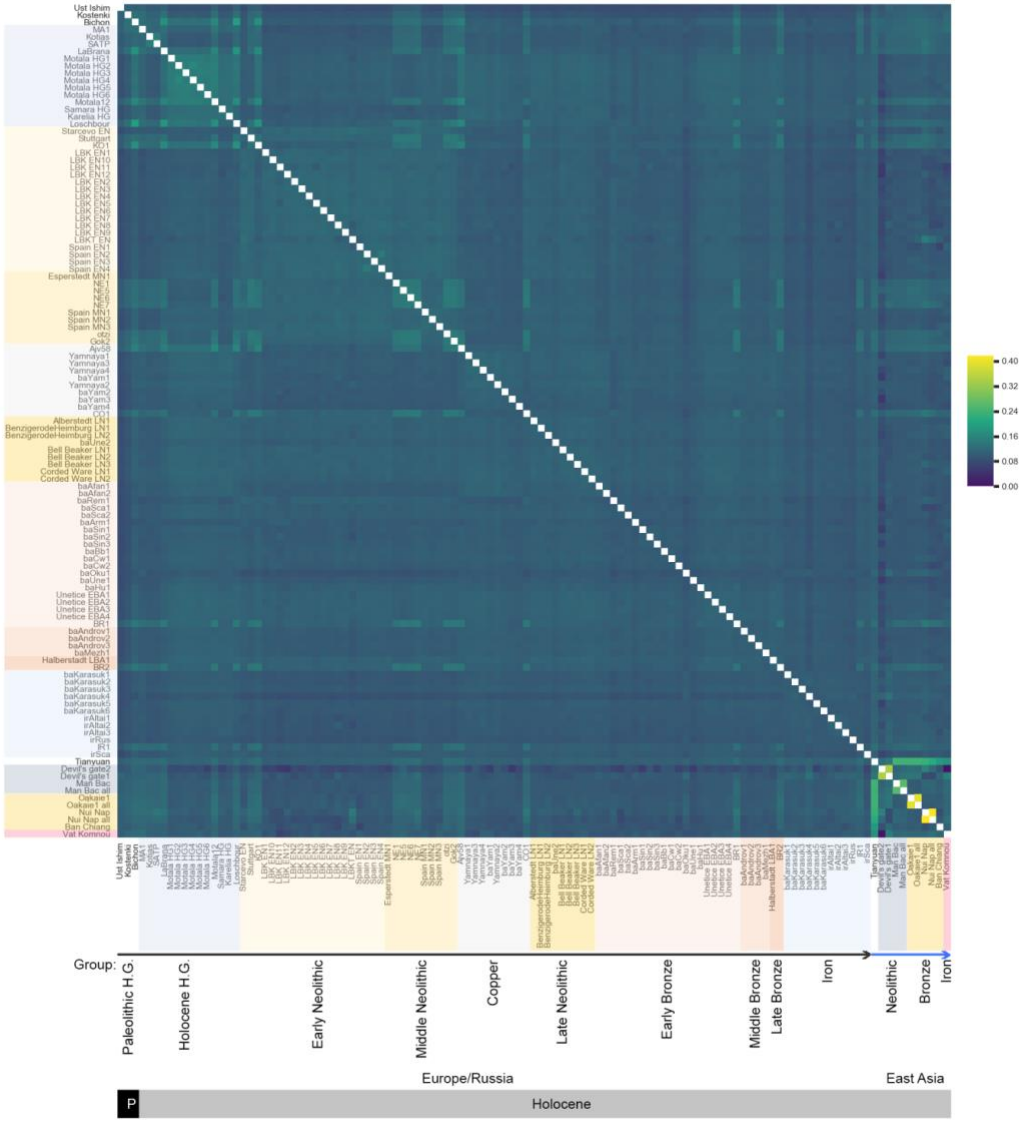


Fig. 42 Genetic cluster of the ancient genomes analyzing with outgroup *f3* statistics

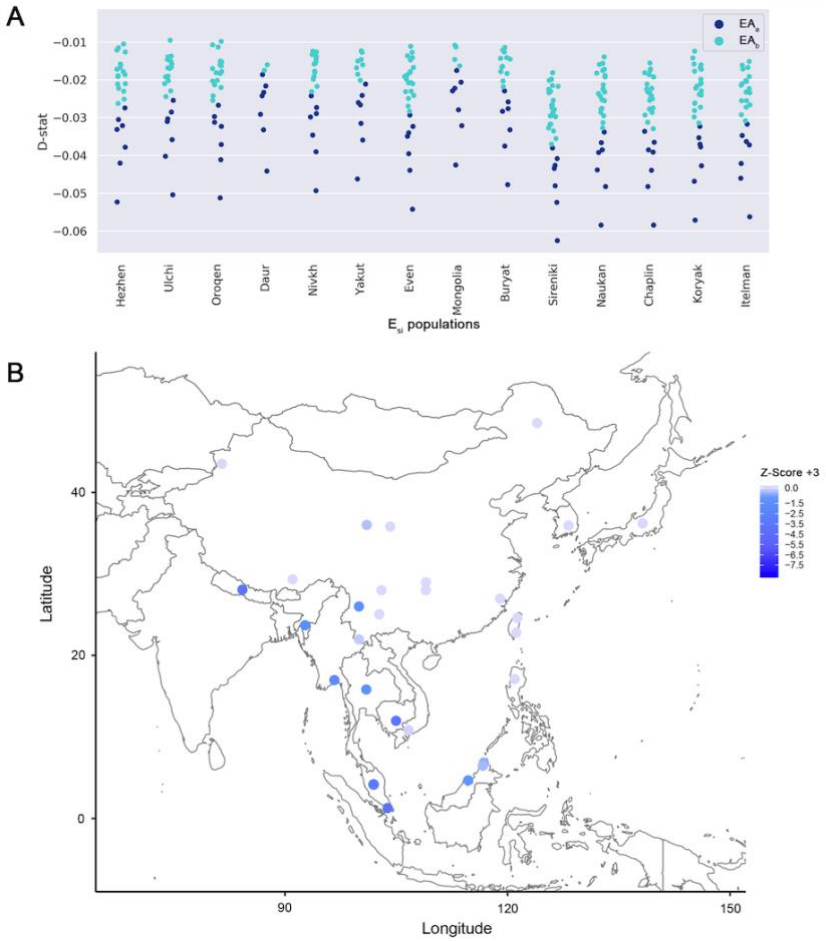


Fig. 43 *D*-statistics with a form of *D*(Yoruba, Tianyuan; E_{si}, $EA_{a/b}$)

(A) The *D*-statistics with a form of *D*(Yoruba, Tianyuan, ancient, present-day) showed that the Tianyuan genome is closely related to E_{si} populations in comparison with $EA_{a/b}$ populations. However, several EA populations (Ami, Atayal, Daur, Hezhen, KOR, Miao, She, Tujia, and Xibo) showed similar levels of divergence from Tianyuan compared to E_{si} populations. (B) To represent gene flow and geographical relationships in Tianyuan, I represented the Z-score of *D*(Yoruba, Tianyuan; Yakut, $EA_{a/b}$). For spot colors, "Z-score +3" values are used to represent "0" for individuals that are not statistically significant.

3.3.3. The ancient gene flow making up the Korean ethnic group

I focused on the gene flow from the Neolithic ancients into the Korean and EA populations. Based on the Tianyuan's gene flow into Neolithic ancients and present-day populations, I hypothesized that either the Neolithic ancient genome contributed to the genetic ancestry of Korean or EA populations independently, or a second gene flow could have occurred (Fig. 40B). First, I investigated gene flow from two Neolithic ancients to Koreans and EA populations, with a form *D*(Yoruba, Devil's Gate/Man Bac, ancient, present-day population). It showed Devil's Gate genomes shared more derived alleles with most of the present-day E_{si} and EA_b populations than with Neolithic Man Bac in Vietnam (Fig. 44A).

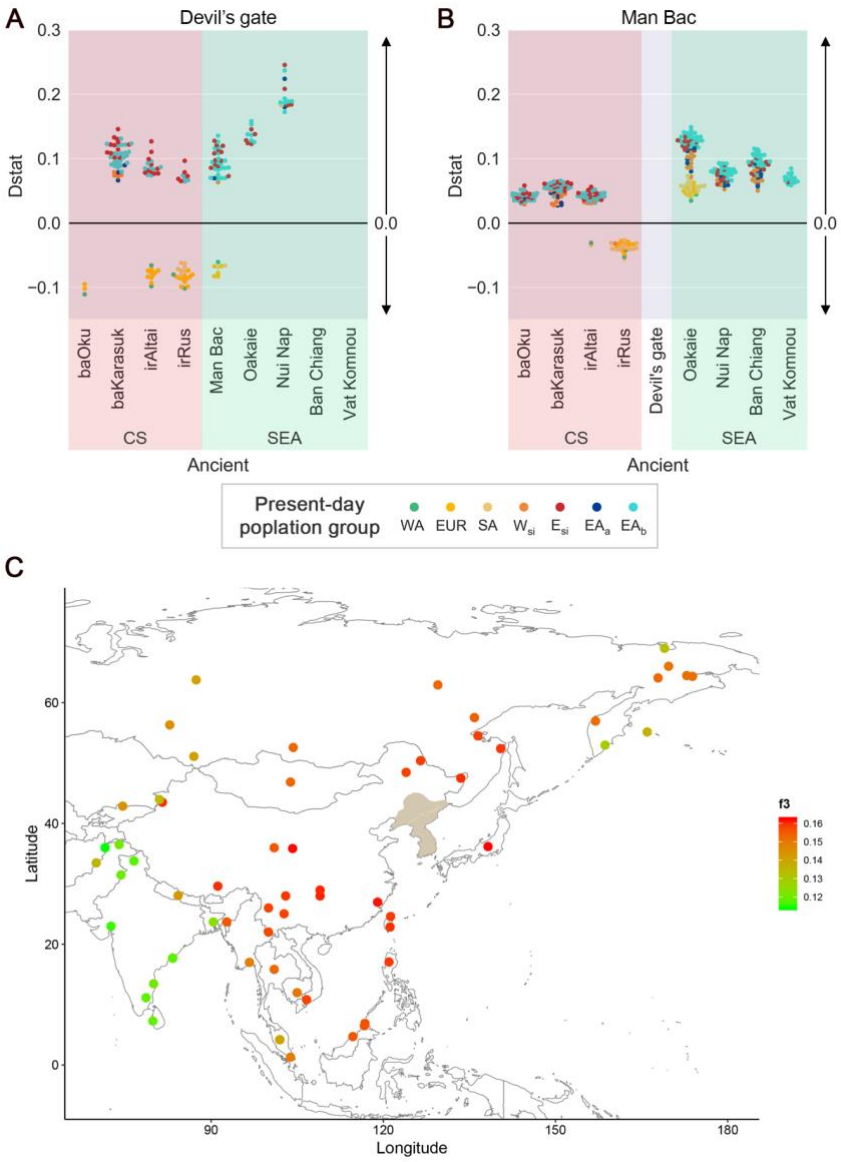


Fig. 44 Bronze and Iron Age gene flows making up the Korean

Ancestry analysis from Neolithic ancients to present-day populations with forms of (A) *D*(Yoruba, Devil's gate, ancient, present-day population), (B) *D*(Yoruba, Man Bac, ancient, present-day population). I represented only the |Z-score| >3 for each *D*-statistics. The positive values represent genetic ancestry to present-day populations and the negative values represent genetic ancestry to ancients at the bottom. The CS represents ancient genomes generated from central steppe regions [92]. (C) Koreans' genetic affinity with neighboring ethnic groups with outgroup *f3* statistics, a form of *f3*(Korean, Y; Yoruba). The spot colors represent the genetic affinity of *f3*-statistics. The overall ancient clustering is represented in Table 4. The predicted historical Korean territories are given in ocher which referenced the website of "About Korea" (http://www.korea.net/AboutKorea/History/Three-Kingdoms-other-States).

Table 4. Outgroup *f3*(Korean, present-day population, Yoruba) analysis

source1	source2	*f3*	std	z	snps	scaledF3
KOR	Japan	0.162178	0.001613	100.525	205320	1
KOR	Han	0.162002	0.001627	99.593	206572	0.99891
KOR	She	0.161644	0.001748	92.458	199957	0.996693
KOR	Miao	0.160831	0.001686	95.389	201578	0.991658
KOR	Tujia	0.160808	0.001648	97.59	201465	0.991515
KOR	Xibo	0.160374	0.001637	97.966	201826	0.988827
KOR	Hezhen	0.160241	0.001655	96.795	201599	0.988004
KOR	Tibet	0.160139	0.001612	99.317	203106	0.987372
KOR	Ami	0.160002	0.001656	96.615	201467	0.986523
KOR	Ulchi	0.159398	0.001689	94.347	201561	0.982783
KOR	Yi	0.1593	0.001631	97.663	201768	0.982176
KOR	Igorot	0.159127	0.001728	92.076	200045	0.981104
KOR	Atayal	0.159022	0.001744	91.172	199924	0.980454
KOR	Oroqen	0.158745	0.001667	95.244	201684	0.978738
KOR	Nivkh	0.158743	0.001683	94.324	201271	0.978726
KOR	Naxi	0.158388	0.001639	96.615	203185	0.976527
KOR	Daur	0.158368	0.001725	91.807	200254	0.976403
KOR	Dai	0.158258	0.001633	96.935	201882	0.975722
KOR	Kin	0.158075	0.001666	94.866	201920	0.974589
KOR	Lahu	0.157578	0.00162	97.253	201820	0.971511
KOR	Rungus	0.156736	0.001711	91.582	199773	0.966296
KOR	Sonsogon	0.156162	0.001789	87.305	199781	0.962741
KOR	Dusun	0.156023	0.001718	90.816	200205	0.96188
KOR	Tu	0.155837	0.001613	96.6	202213	0.960728
KOR	Mizoram	0.155596	0.001689	92.125	199773	0.959235
KOR	Even	0.154337	0.001588	97.206	203871	0.951438
KOR	Mongolia	0.152968	0.001632	93.742	202300	0.942959
KOR	Buryat	0.152905	0.001653	92.498	202556	0.942569
KOR	Koryak	0.152769	0.001624	94.057	201609	0.941727
KOR	Yakut	0.152695	0.001667	91.595	202377	0.941269
KOR	Thailand	0.152643	0.001615	94.505	202015	0.940947
KOR	Itelman	0.152176	0.001714	88.773	200399	0.938054
KOR	Sireniki	0.151959	0.001651	92.061	200502	0.93671
KOR	Chaplin	0.151173	0.001675	90.254	200396	0.931842
KOR	Naukan	0.150626	0.001655	91.005	201961	0.928455
KOR	Cambodian	0.149918	0.001631	91.906	202736	0.92407
KOR	Cosmomalay	0.149298	0.001641	90.989	202918	0.92023
KOR	Burmese	0.148392	0.001684	88.132	200757	0.914619
KOR	Orseletar	0.147134	0.001718	85.643	201439	0.906828
KOR	Altai	0.144894	0.001655	87.533	203472	0.892954
KOR	Kyrgyz	0.144132	0.001587	90.795	203773	0.888235
KOR	Nemendriq	0.143471	0.001632	87.933	202543	0.884141
KOR	Kusunda	0.14287	0.001678	85.122	200719	0.880419
KOR	Ket	0.140603	0.001657	84.85	203675	0.866379

KOR	Tubalar	0.14053	0.001623	86.579	203943	0.865927
KOR	Nejehai	0.139637	0.001645	84.894	202294	0.860396
KOR	Aleut	0.13701	0.001594	85.951	204376	0.844126
KOR	Hazara	0.136146	0.001757	77.498	201674	0.838775
KOR	Uygur	0.133809	0.001603	83.467	204929	0.824301
KOR	Chukchi	0.132985	0.001579	84.22	202063	0.819198
KOR	Tlingit	0.128466	0.001544	83.225	205458	0.79121
KOR	Relli	0.124113	0.001665	74.534	202115	0.764251
KOR	Bengali	0.123123	0.001514	81.35	205579	0.758119
KOR	Madiga	0.122638	0.001609	76.212	202325	0.755116
KOR	Kapu	0.120908	0.001509	80.126	205559	0.744401
KOR	Mala	0.120762	0.001548	78.032	205244	0.743497
KOR	Irula	0.120616	0.001508	79.976	205248	0.742593
KOR	Burusho	0.120046	0.001583	75.818	202665	0.739063
KOR	Punjabi	0.119759	0.001492	80.253	208089	0.737285
KOR	Sinhalaese	0.118915	0.001576	75.442	202340	0.732058
KOR	Tamil	0.11804	0.001481	79.68	205846	0.726639
KOR	Brahmin	0.117923	0.001542	76.468	206021	0.725914
KOR	Kashmir	0.117523	0.001606	73.17	202820	0.723437
KOR	Gujarat	0.115603	0.001511	76.521	206084	0.711546
KOR	Kalash	0.114026	0.00148	77.021	206044	0.701779

From the Devil's Gate genome near North Korea, I observed these present-day populations are equivalent to the genetic relationship with Ban Chiang and Vat Komnou ancients who are ancestors of Austroasiatic speakers [79]. In addition, I observed local genetic transitions from Oakaie (Late Neolithic and Bronze Age in Myanmar) and Nui Nap (Bronze Age in Vietnam) to EA populations. Several Esi and EAb populations, such as Korean, Japanese and several Chinese (Hezen, and She), and Russian (Ulchi) ethnic group, still had dominant genetic contributions from Devil's Gate compared with Oakaie and Nui Nap ancients. This suggests that local genetic differences observed in present-day $EA_{a/b}$ populations (Fig. 38C) were influenced by a new genetic influx from the Bronze Age to Iron Age in Southeast Asia. I also observed *D*(Yoruba, Devil's gate, baOku, present-day E_{si} or E_{Ab}) ~0 (Fig. 44A) and *D*(Yoruba, baOku, E_{si}, EA_b) ~0. According to these statistics, the baOku genomes are equally closely related to present-day E_{si} and EA_b populations, which is different from the dominant ancestry of the E_{si} populations in baKarasuk (Iron Age in Russia) and irAltai (Iron Age in Russia). Unlike the Devil's Gate's ancestry, the Neolithic Man Bac shares more derived alleles with most of the present-day E_{si} and EA_b populations than either the Bronze Age ancSEAs (Oakaie, Nui Nap, Ban Chiang) or ancCSs (baOku, baKarasuk, irAltai) (Fig. 44B). This suggests the Neolithic Man Bac is the basal ancestry for the present-day E_{si} and EA_b populations. No genetic drift was observed from Neolithic Man Bac to Devil's Gate ancient and present-day populations (Fig. 44B). I also analyzed genetic associations of ancCS to other ancients and present-day populations with a form of *D*(Yoruba, ancCS; ancient,

present-day populations) (Fig. 45). It inferred that present-day E_{si} and EA populations and ancSEA are equally related to ancCS by sharing similar levels of ancCS-derived alleles. It is an agreement with genetic admixture patterns of Asian ancestry in CS ancients[93, 94]. It supports genetic admixture between ancCS and present-day EA populations, however, it cannot explain how and how many events the ancCS influence toward EA occurred. I also observed the first evidence of the genetic divergence of Vat Komnou and several EA_b (Southeast Asian and Southern China) populations from Man Bac (Fig. 44B). This supports the idea that these ancients are new genetic resources that genetically influenced EA (Fig. 40A). I observed several possible ancient founders by *D*-statistics, however, it could not clearly resolve the current genetic makeup of Korean. To resolve the genetic relationship of the genetic makeup of Korean, I additionally analyzed the admixture pattern of the ancient/present-day Southeast Asians and Devil's Gate ancients to Koreans with admixture *f3* statistics (Table 5). Notably, the combinations of the Devil's Gate genome and ancSEAs better represent the current Koreans than those of Devil's Gate and modern Southeast Asians. Specifically, I observed the lowest admixture *f3*-statistics when source 1 was Vat Komnou (Iron Age in Cambodia), followed by Nui Nap (Bronze Age in Vietnam). In a previous study, Nui Nap was a new genetic component close to present-day Vietnamese and Dai but not the ancestors of Austroasiatic speakers[79]. Meanwhile, next ancSEAs with lowest admixture *f3*-statistics were Ban Chiang and Man Bac who are also ancients of Austroasiatic speakers. In order to investigate whether the ancSEA genetic components migrated into Korea, I analyzed the Koreans' genetic affinity with present-day populations by outgroup *f3*-statistics with a form of *f3*(Korean, present-day populations; Yoruba) (Fig. 44C and Table 4). It showed the group with the highest genetic affinity with the Koreans were the Japanese. The southern Chinese (Han, and She) had a higher genetic affinity with Koreans than the present-day Lau or Vietnamese, which is consistent with the admixture results (Fig. 38C). This suggests that the genetic components of South Chinese were transferred into Korea after admixing with Vat Komnou and Nui Nap ancestries (Fig. 44C). These lines of evidence support the conclusion that populations who carried Devil's Gate and Man Bac genomes admixed throughout the EA_b and E_{si} regions until the Neolithic period, probably accompanied by the climate changes and barriers. After the Bronze Age, the admixed genetic ancestry of the Vat Komnou and Nui Nap migrated to Korea due to rapid cultural and technological advances.

Table 5. Admixture *f3* statistics[a]

Source1	Source2	Avg. *f3*	Min. *f3*	Max. *f3*
Vat Komnou	Devil's gate2	-0.192366	-0.22219	-0.173976
Nui_Nap	Devil's gate1	-0.13199	-0.13199	-0.13199
Ban_Chiang_all	Devil's gate1	-0.127784	-0.127784	-0.127784
Ban_Chiang	Devil's gate2	-0.118145	-0.118145	-0.118145
Nui_Nap_all	Devil's gate1	-0.10339	-0.10339	-0.10339
Man_Bac	Devil's gate2	-0.055678	-0.056621	-0.054339
Atayal_EA	Devil's gate2	-0.038359	-0.04107	-0.035966
Ami_EA	Devil's gate2	-0.0380293	-0.040296	-0.036663
Lahu_EA	Devil's gate2	-0.036503	-0.039709	-0.034341
Kinh_EA	Devil's gate2	-0.034616	-0.036383	-0.031549
Thai_EA	Devil's gate2	-0.0334685	-0.035207	-0.03173
Dai_EA	Devil's gate2	-0.032952	-0.033388	-0.032296
Cambodian_EA	Devil's gate2	-0.032376	-0.032407	-0.032345
Tujia_EA	Devil's gate2	-0.0314865	-0.032745	-0.030228
Han_EA	Devil's gate2	-0.030894	-0.031301	-0.030493
She_EA	Devil's gate2	-0.0303735	-0.031006	-0.029741
Miao_EA	Devil's gate2	-0.03032	-0.03032	-0.03032
Yi_EA	Devil's gate2	-0.030312	-0.030312	-0.030312

[a] the notation of admixture *f3* statistic: *f3*(Source1, Source2; KOR) and only represented with |Z-score| > 3

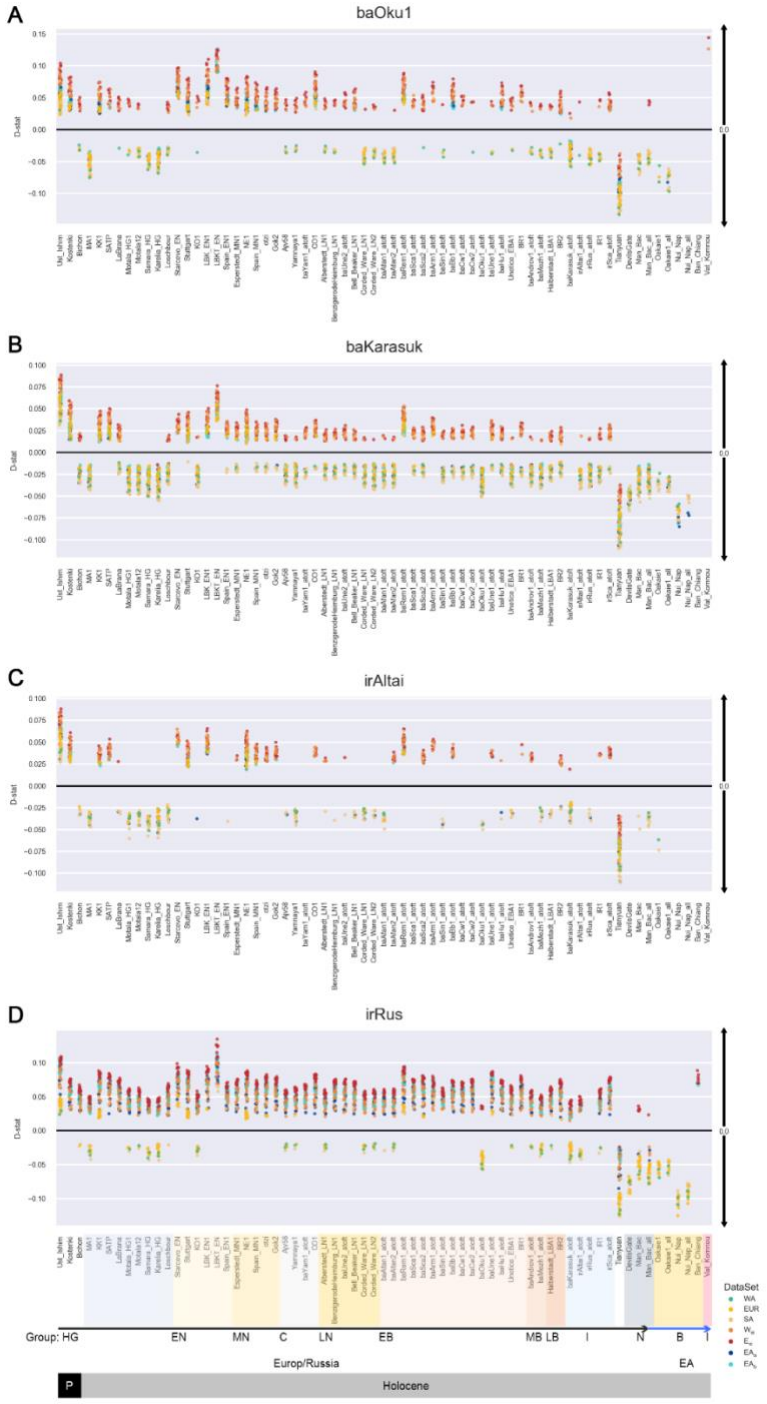


Fig. 45. *D*-statistics with a form of *D*(Yoruba, ancCS; ancient, present-day)

D(Yoruba, ancCS, ancient, present-day) test suggests ancCS genomes are genetically closer to present-day E_{si} populations than European ancients. It also inferred that present-day E_{si} and EA populations and ancSEA are equally close to ancCS, relatively. I represented only the |Z-score| >3 for each D-statistics. The positive values represent genetic ancestry to present-day populations and the negative values represent genetic ancestry to ancients at the bottom.

3.3.4. Korean haplotype analysis reveals multiwaves of genetic components

I analyzed haplotype distributions using WGS data of 88 unrelated Koreans generated from the KoVariome database[15]. Nonrecombining Y-chromosome analysis showed a significant proportion of the "O" haplogroup in 55 male Koreans, 29% "O2b" and 42% "O3" (Fig. 46A). The next most frequent Y-chromosome haplogroup was "C" (18%). The Y-chromosome haplogroup distribution agreed with well-established Y-chromosome haplogroup "O" expansion and colonization within the Korean Peninsula[72] . A comparison with the global Y-chromosome haplogroup distribution suggested that haplotype "C" is widespread in Siberia, whereas "O" haplogroups show a spatial distribution in Southeast Asia[95, 96]. This strongly suggests a dual origin for Korean males. In contrast to the Y-chromosome distribution, mtDNA haplotypes reflect a more complex genetic history (Fig. 46B). The most frequent mtDNA haplotype was "D" (34%) and ten additional mtDNA haplogroups ("M," "B," "N," "G," "F," "R," "A," "C," "Y," and "Z") were identified with frequencies ranging from 23% to 2%. I constructed an mtDNA tree combining 11 ancients, and 99 present-day $EA_{a/b}$ and Siberian (E_{si} and W_{si}) mtDNAs (Fig. 46C). Similar to the global human-mtDNA phylogeny, the mtDNA tree shows two major clades, M′ and R′, dominantly distributed in EA populations[97]. It also shows two mtDNA dispersions ~40 and 20 ka, which account for 62% and 38% of the present-day Koreans, respectively. The earlier dispersed mtDNAs included "N/Y/A," "D," and "B/R" which were distributed to 16%, 34%, and 12% of Koreans, respectively. The mtDNA haplotypes of the "N/Y/A" and "D" were clades co-clustered with present-day Siberians as well as the Devil's Gate ancients, representing Eurasian ancestry. The "A" haplogroup was also frequently observed in the early and middle Bronze Age Okunevo peoples[79], who were culturally associated with baKarasuk [79]. I also identified ancient mtDNA "R′" divergent into "B/R," accounting for 12% of Koreans, that also expanded ~40 ka. The root of this clade was Tianyuan, and also co-clustered with Vat Komnou ancients and present-day Chinese, representing EA ancestry. This could explain the genetic influence of the Tianyuan on Korean genomes via ancSEA. These old mtDNA waves accounted for human migration in the late Pleistocene when the Yellow sea of Korea was land, therefore, the west coast of Korea was connected to the mainland of China. The later dispersed mtDNA haplogroups consisted of "G/C/Z," "M," and "F" which account for 19%, 12%, and 7% of Koreans, respectively. The "G/C/Z" clades co-clustered with Siberians and Bronze Age Nui Nap in Vietnam. However, the genetic origin of the Nui Nap is still unknown. On the other hand, the mtDNA haplogroup "C" is frequently observed from the early and middle Bronze Age Okunevo peoples who lived in central steppe regions[79]. The mtDNA topology and haplotype frequency in Okunevo imply a genetic association between Nui Nap and central steppe ancients. Both of the "M" and "F" clades showed subsequent diversification from ancient mtDNA haplogroups of ancM (M′) ~20 ka and ancR (R′) divergent in 60 ka, respectively. These clades explain southern waves of human migration by co-clustering with EAb populations. In particular, two ancients of Austroasiatic speakers, Man Bac and Ban Chiang, co-clustered in the mtDNA "M" lineage (Fig. 46C). It suggests that a subsequent expansion of this clade

can be associated with the expansion of the Austroasiatic speaking population[79]. Haplotype analysis and the phylogenetic tree of the mtDNA support a continuous genetic influence from the north and south into Korea.

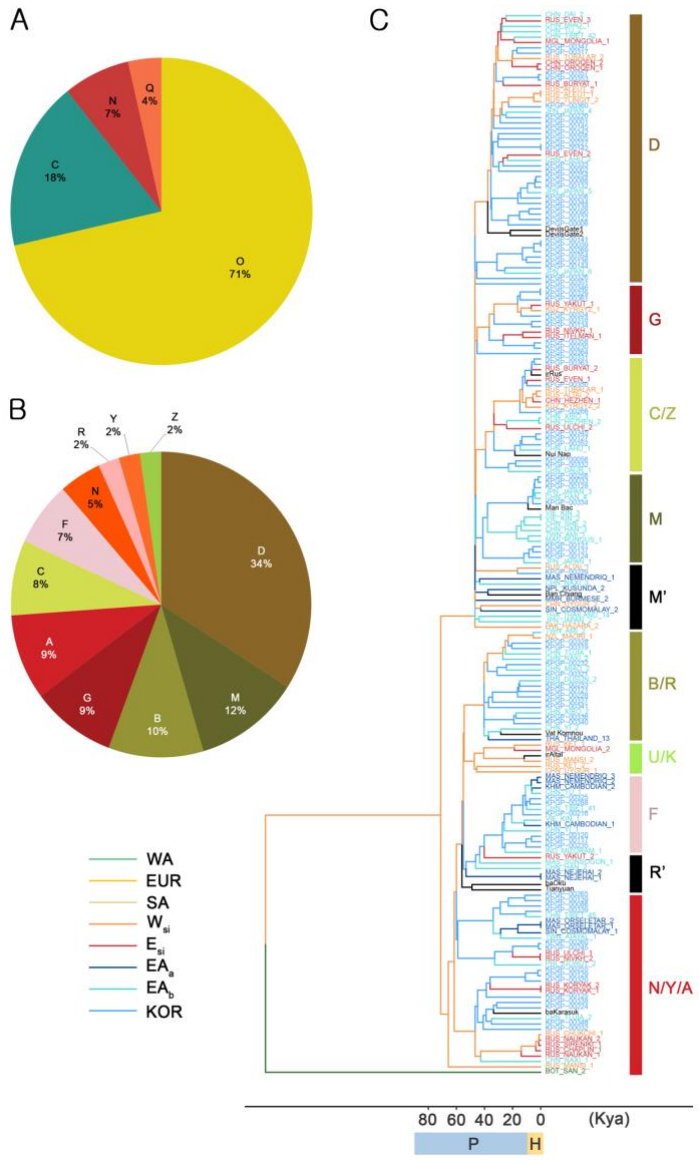


Fig. 46 Haplotype distribution in the Korean population

(A) Y-chromosome haplotypes from 55 male Koreans, (B) mtDNA haplotypes in 88 Koreans, (C) A phylogenetic tree of mtDNA haplotypes constructed using the neighbor-joining method with bootstrap=1,000. I give the dominant mtDNA haplogroup clusters on the right of the tree. The ancient haplogroup is represented by M' and R'. Abbreviations: P: Pleistocene, and H: Holocene.

3.3.5. Admixture time estimation for Koreans

I estimated the admixture time of Koreans using 286,222 SNPs and obtained significant prediction results from only three populations as references; Yakut, Han, and Japanese (Table 6).

Table 6. Estimation of admixture date of Koreans

Population group	Reference population	No. of sample	Admixture Time [a]		Z-score	*P*-value
			Generation	Years		
E_{si}	Yakut	20	189.05 (65.86 – 312.24)	5,482 (1910 – 9055)	3.01	$1.3 \cdot 10^{-3}$
EA_b	Han	33	123.56 (72.05 – 175.07)	3,583 (2089 – 5077)	3.85	$5.9 \cdot 10^{-5}$
EA_b	Japanese	29	97.47 (34.60 – 160.35)	2,827 (1003 – 4650)	3.71	$1.0 \cdot 10^{-4}$

[a]The admixture time is shown in generations before the present. The number in the parentheses indicates 95% confidence interval of the generation and years.

The estimated admixture time was 5,482, 3,583, and 2,827 YA when I used the Koreans itself as one reference and Yakut, Han, and Japanese as the other comparison reference population, respectively. The estimated admixture time with Japanese (97 generations away from the Japanese) is slightly earlier than the admixture date of the mainland Japanese (52 generations) estimated by Takeuchi et al.[74]. I summarized the model of the genetic influence by pre-Neolithic Tianyuan to Iron Age Vat Komnou on Koreans in Fig. 47. This model supported the above gene flows well, suggesting Koreans contain prehistoric genetic components derived from Devil's Gate and Man Bac groups both of whom are divergent from Tianyuan ancestry. The Neolithic Man Bac genome dominantly inherited the genetic components of Tianyuan and showed its genetic components widely distributed in EA. However, the Bronze and Iron Age ancients, such as Oakaie, Nui Nap, and Vat Komnou, seem to have much altered genetic components of EA_b genomes (70%). This is consistent with the EA_b ancestry frequency in contemporary Koreans. This model generally describes well the gene flow among the three Northeast Asians; Korean, Chinese, and Japanese.

3.4. Conclusion

I analyzed the haplotype distributions of 88 Koreans compared with ancient and modern whole genomes and suggested two major haplotype expansion events. A comprehensive genome comparison confirmed that Koreans possess dual ancestral genetic components originating broadly from East Siberia (E_{si}) and East Asia (EA_b). Ancient genome comparisons revealed that the genetic makeup of Koreans can be best described as an admixture of the Neolithic Devil's Gate genome in Russia and the Iron Age Vat Komnou in Southeast Asia. The analyses of ancient and present-day populations suggest a long and gradual admixture model of two Neolithic founders, the Devil's Gate founder in Russia and the founder from Tianyuan Cave in China. These two major components were admixing throughout East Siberia and East Asia for an extended time up until the Neolithic period. Subpopulations of current East Asians, as well as modern Koreans, were probably established by a later regional genetic transition during the Bronze Age. The peopling of Korea is most likely a part of large population expansion and the subsequent admixture events which occurred in East Asia, rather than a unique isolated event or migration. I think that this kind of recent rapid expansion and admixture could be general models for other East Asian and Southeast Asian populations in which Bronze and Iron Age populations expanded and admixed with other peripheral region populations.

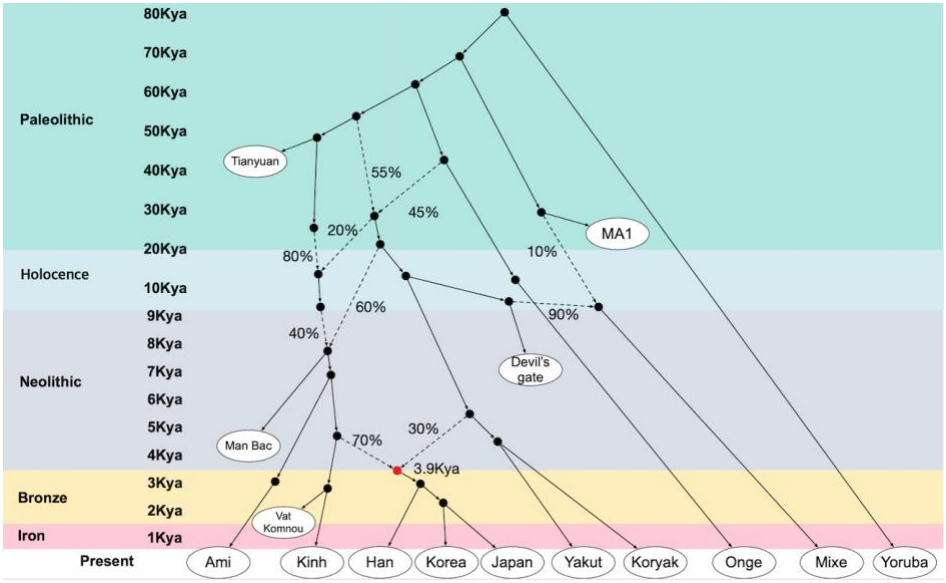


Fig. 47 Admixture tree model depicting the historical genetic makeup of Korean

A qpgraph [88] fitted on an admixture model depicting the historical genetic makeup of Koreans and other Asians. I fitted the admixture tree model with ancient genomes associated with EA_b populations to make a model that could best explain the gene flow that makes up Koreans and hence the admixture model information for E_{si} ancestry has been simplified. Based on the *D*- and *f3* statistics and previous reports [79], I set the skeletal tree (Fig. 48A) and extended the model by adding ancient and present-day

individuals (Fig. 48). The average admixture time of Koreans is noted next to the red circle which was estimated by ALDER (Table 6). Black circles represent ghost genomes in ancestral genetic lineages lacking any evidence for a time calibration and new groups may be added when more ancient populations are found and sequenced. Black lines represent the gene flow and dotted lines represent admixture events with the marked proportions estimated by qpgraph analysis.

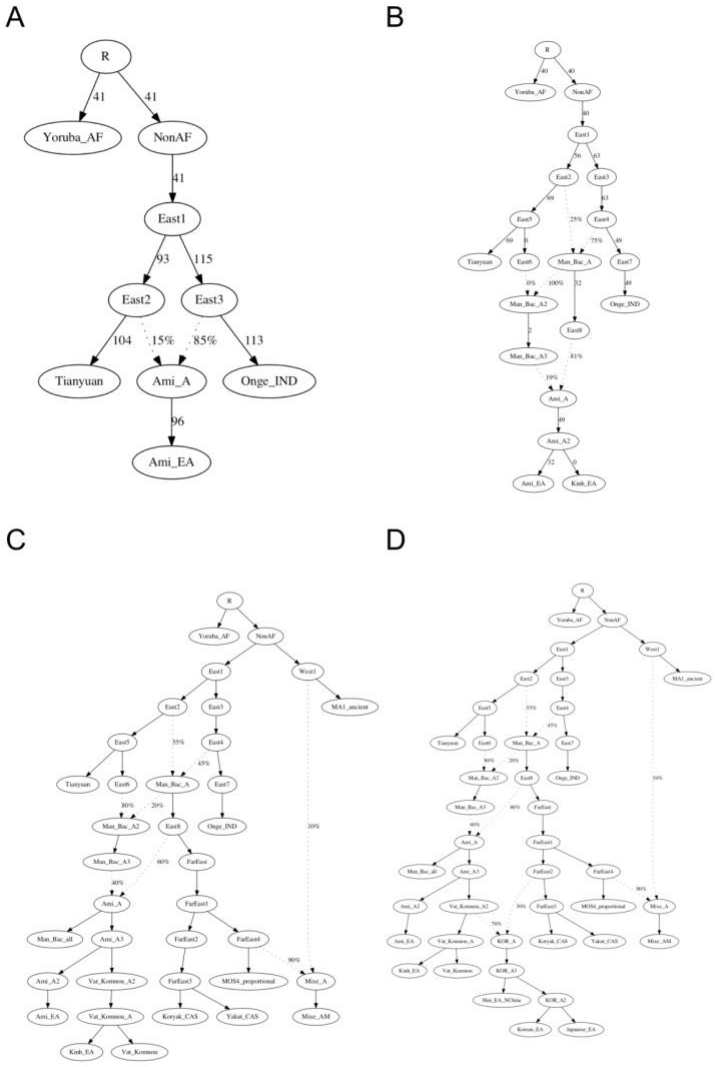


Fig. 48. Four tested admixture tree models by qpgraph.

(**A**) Admixture tree model skeleton adapted by previous reports [79] (worst-fitting Z=0.044). (**B**) Admixture tree model after adding Kinh from the first model (worst-fitting Z=-3.887). (**C**) after adding Far East samples (Devil's gate, Ulchi, Koryak, Mixe, and MA1) (worst-fitting Z=3.317). (**D**) Final admixture tree model by adding the Korean, Han, and Japanese populations (worst-fitting Z=-3.686)

IV. Chapter 3. Korean Genome Project portal

4.1. Introduction

Genome data is one of the largest bigdata that includes various biological information. High-throughput sequencing technology has generated a massive amount of sequencing data from various organisms on Earth. Storing and distributing these huge data to the public can elevate the synergetic effect of inter- or intraspecies comparative genomics. The National Center for Biotechnology Information (NCBI) is a major resource for sharing and distributing genome sequences and additional information[98]. The University of California Santa Cruz (UCSC) genome browser is a handy web application for visualizing and analyzing genomic resources, especially while searching for various genomic annotations on target regions via a web browser[99]. The Ensembl genome browser provides access to information on the reference genome sequences and genomic annotations of many vertebrates[100]. It also supports studies in comparative genomics, evolution, variations, and transcriptional regulation by providing gene annotation, computed multiple sequence alignment, etc.

Specifically, with respect to human population genomics, the Exome Aggregation Consortium (ExAC) browser presents qualitative and quantitative information of the genes and variants in the ExAC dataset[101]. This browser now supports the gnomAD dataset, which covers wider genomic regions and more genomes from across the world. Recently, the PGG.SNV database was developed which presents 265 million SNVs collected from 220,147 present-day genome and 1,018 ancient genomes across 977 global populations[102]. PGG.SNV provides information on worldwide distribution, natural selection pressure, and genomic diversity, which can be beneficial resources for human population comparative genomic studies. For Korean genomes, the Korean National Institute of Health has developed the Korean Reference Genome Database (KRGDB) browser[103]. Although the KRGDB is a large-scale variant database of 1,722 Koreans, I was unable to access the information since the browser link provided in the paper was not working at the time of writing this dissertation.

This chapter presents the Korean Genome Project Portal (KGPP), which provides the variants and phenotype-association information via a web application. The KGPP includes a search engine for finding the variants and their annotations on a target gene. It also provides the associated traits and statistical association metrics of each variant. The chapter also presents an open API system, developed by me, that shares the allele frequencies and associations data freely via a URL and enhances the developments of other applications that use Korea1K data for developers and researchers.

4.2. Methods

4.2.1. Data sources

The Korea1K variome was used for the KGPP. As described in the Korea1K paper, variants were annotated using the VEP ver. 95[33]. I used the genome-wide association study (GWAS) results of Korea1K to show the identified associations with phenotypes of each variant. The KGPP currently contains information on 41,977,415 variants information, and 2,547 associations with quantitative traits.

4.2.2. Design of the Korean Genome Project portal

To enable effective access to the Korea1K variome efficiently through a web application, I developed the Korean Genome Project Portal (KGPP), which consists of three major parts: the client-side, the server-side API, and the database management system (Fig. 49).

Client-side

The ReactJS provides a client-side web page that renders images on the user's browser (Fig. 49A). The Bootstrap4 framework was used to make a responsive HTML backbone. PlotlyJS was used to visualize the allele frequency distributions as a responsive plot.

Server-side API and database management system

The Go programming language and the ECHO framework provides the Korea1K data in the NoSQL DB to the client-side web page in the JavaScript Object Natation (JSON) format through the API (Fig. 49B). The information on the 41,977,415 variants information and the 2,547 associations with quantitative traits was stored in three collections in the MongoDB NoSQL database (Fig. 49C).

Optimization

I optimized the database by indexing the MongoDB collection "variants" by variant ID and gene symbol, thereby allowing me to search for the variants of a specific target gene from 41 million variants.

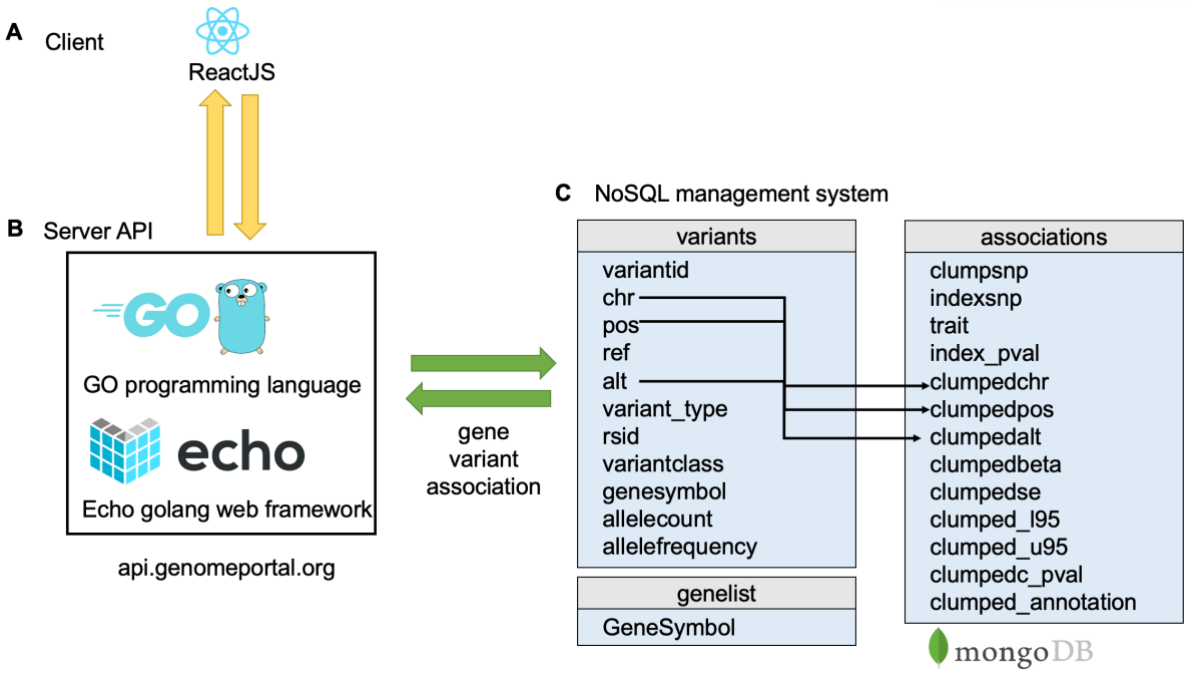


Fig. 49 Architecture of the Korean Genome Project portal. A) Client-side developed by ReactJS v16.13.1. B) Server-side API developed using the Go programming language and the ECHO web framework. C) NoSQL management system by MongoDB and database schema.

4.3. The Korean Genome Project Portal

The KGPP has three main web pages differentiated by the visualization function. The homepage of the KGPP has a search engine that enables search by a gene symbol (Fig. 50). The web page for each gene shows the allele frequency distribution on the gene, with different colored bars representing the variant functional classification, and a list of the variants with links to the corresponding variant pages (Fig. 52). The variant pages include variant annotation information (e.g., genomic position, allele frequency, and functional classification) and associated clinical traits from the Korea1K study (Fig, 53).

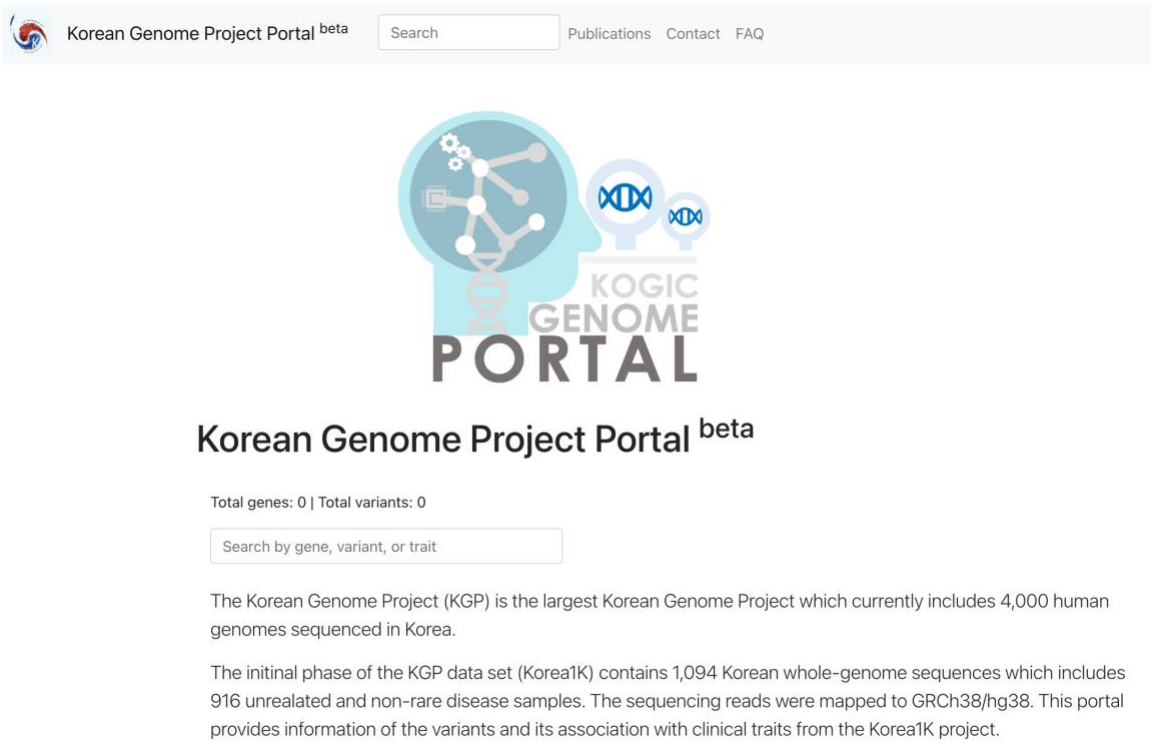


Fig. 50 Homepage of the Korean Genome Project portal.

4.3.1. The Korean Genome Project portal API

I developed an API for retrieving the Korea1K data stored in the database and for sharing it with the client-side web application of the KGPP ("api.genomeportal.org"). The API has three major routes, each of which returns the array of variant, gene, or phenotype-association information in JSON format (Fig. 51). As a result of indexing the database with variant IDs and gene symbols, the time taken in searching by variants or gene has been dramatically decreased.

Fig. 51 Sample result of an API request. A list of *LPA* gene variants is returned as an array of JavaScript objects in the JSON format.

4.3.2. Gene page

The KGPP gene page shows the list of variants for a target gene. The page starts with the metadata of the gene and the count of variants based on the functional location of the variant. Along with the basic information about the gene and its variant count, the page shows the allele frequency distribution across the genomics position (Fig. 52). The functional location of the variants is represented by a different color. The allele frequency distribution plot can be zoomed in and out of or saved as an image file. Below the plot is a table containing along with their page numbers, a list of the variants in the Korea1K database with the variant ID, genomic position, dbSNP rsID, and allele frequency. Users can filter the variants using the multiple filtering options. The first column of the table contains links to the corresponding variant pages.

Allele frequency information

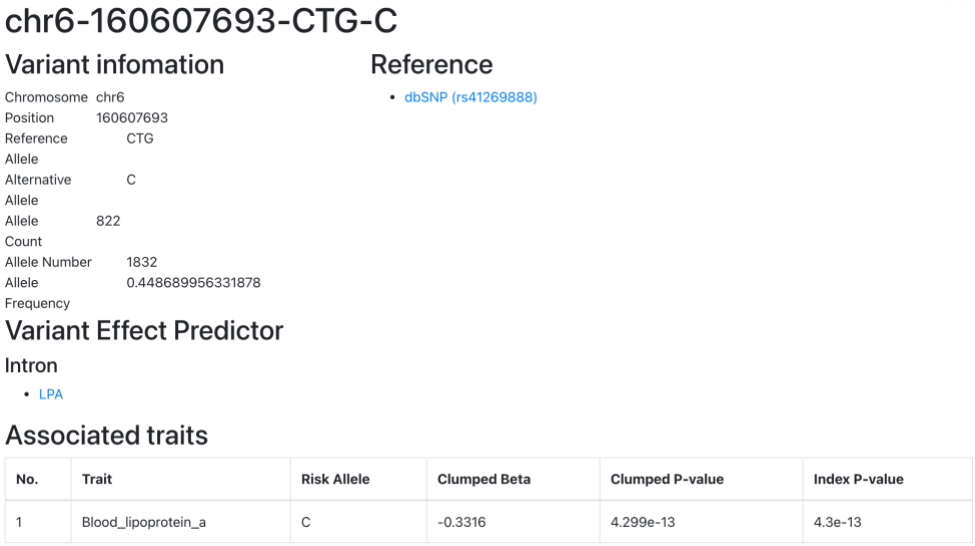
<< < > >> Page **1 of 106** | Go to page: 1 Show 20

ID	Chromosome	Position	Ref	Alt	Variant type	rsID
Search 2104 records...	All	Search 2104 records...	Search 2104 records...	Search 2104 records...	All	Search 2104 records..
Search: 2104 records...						
chr6-160526817-T-C	chr6	160526817	T	C	SNP	novel
chr6-160526870-A-G	chr6	160526870	A	G	SNP	rs190047149
chr6-160526878-G-T	chr6	160526878	G	T	SNP	rs1016787109
chr6-160526903-T-G	chr6	160526903	T	G	SNP	novel
chr6-160526910-A-T	chr6	160526910	A	T	SNP	novel
chr6-160526985-G-A	chr6	160526985	G	A	SNP	rs151239228
chr6-160526986-C-A	chr6	160526986	C	A	SNP	rs79977623
chr6-160527104-A-C	chr6	160527104	A	C	SNP	rs12664092
chr6-160527160-C-G	chr6	160527160	C	G	SNP	novel

Fig. 52 Sample gene page of the Korean Genome Project portal. Information on allele frequency distribution across the *LPA* gene and the list of variants found in *LPA* is retrieved from the Korea1K variome.

4.3.3. Variant page

The KGPP variant page shows detailed information on the variants in the Korea1K. The page first displays metadata of the variant, for example, genomic position, allele count, and frequency. If the variant has been reported in the dbSNP, a direct link to the corresponding dbSNP web page is provided on the right-hand side of the page. Below the variant information, the page shows annotations of the variants along with their gene symbols. If the variant is located on multiple genes, the variant page displays all possible gene symbols and functional location annotations.



chr6-160607693-CTG-C

Variant infomation

Chromosome	chr6
Position	160607693
Reference Allele	CTG
Alternative Allele	C
Allele Count	822
Allele Number	1832
Allele Frequency	0.448689956331878

Reference

- dbSNP (rs41269888)

Variant Effect Predictor

Intron

- LPA

Associated traits

No.	Trait	Risk Allele	Clumped Beta	Clumped P-value	Index P-value
1	Blood_lipoprotein_a	C	-0.3316	4.299e-13	4.3e-13

Fig. 53 Sample variant page of the Korean Genome Project portal. Information of a variant on *LPA* and its association with the clinical trait(s) is shown.

4.4. Discussion

This chapter introduces the Korean Genome Project portal, which was developed to share information on allele frequency and GWAS variants from the Korea1K database, with the public through a web application as efficiently as possible. An open API system was also designed to enable researchers to search the database for desired data easily via a simple URL and web communication. Though this portal acts as an efficient searching and visualization system, there is a lot of room for improvement. First, the search engine currently on supports gene symbols, and optimization of the database indexing and search engine are needed to execute a search by variants or by other keywords. Second, there is a lack of information on allele frequencies in different ethnic groups, and public databases like the GWAS catalog in the KGPP. This additional information can act as informative resources for comparative analyses and evolutionary studies. Finally, sharing the individual data, including information on sequencing read mapping status and individual genotypes, is nearly impossible due to Korean laws and regulations. Freely available individual genotypes and matching clinical information should be made widely available as reference data for use in imputation and GWAS, and for the development of a platforms for personalized medicine. Nevertheless, the concept behind this genome portal system can be extrapolated for use as a personalized portal system that provides personal genome and variome annotations, including those from the Korea1K variome. The portal can be used to provide users with personalized risk information or genomic feature predictions based on the genome sequences that they upload in the portal. This information may enable people to adjust their lifestyles which can help to prevent or delay disease occurrence.

V. Conclusion

In this dissertation, I firstly presented, as the large-scale Korean population genome data, a comprehensive WGS analysis of 1,094 Koreans (Korea1K). Korea1K showed that the Korean population is genetically homogeneous compared with other human populations. Using the whole-genome level of variome and matched information of 79 quantitative clinical traits, I found nine GWAS index variants that have not been previously reported, though their clumped loci have. This suggests that a large-scale variome from WGS can, in fact, identify more significantly associated loci with the traits, by covering wider genomic positions for the target loci in GWAS.

Korea1K also showed that a large-scale single population variome can increase the power of genotype imputation to be used as a reference panel. However, despite the large amount of population genomic data available, I was unable to use CNV and TE analyses using short-read sequencing to fully identify unique or Korean-specific genomic features in terms of phenotype association or genomic selection. This is a possible limitation of the short-read sequencing. Personal *de novo* genome assemblies (PRG: Personal Reference Genomes), using a combination of long-read sequencing technology and other genome sequencing technologies like Hi-C, may be able to identify the structural variations covering longer variants with higher accuracy, since the assemblies have longer sequence contiguity.

Another application of the population-scale Korean WGS is in investigating the origin of Koreans. Due to recent advancement in the paleogenomics, I was able to collect many ancient and modern human variome resources and build a model of the admixture of East Asians and the genetic origin of Koreans. The model suggests that the admixture events initially occurred mainly outside the Korean peninsula and continuous spread and localization in Korea, which is consistent with the general admixture trend of the East Asian region in the Bronze-Iron age. However, the dataset still does not include ancient genomes from the Korean peninsula. If multiple time points of ancient genomes inside the Korean peninsula become available, this constructed model can have a higher and more accurate resolution and even provide a detailed inter-country model among Chinese, Japanese, and Koreans.

The Korean Genome Project portal that I have presented in this dissertation provides Korea1K variant and phenotype-associated information via a public web application. This portal can be used for many future clinical (especially, relating to rare diseases) and population studies by providing the allele frequencies of the variants identified. The Korean Genome Project portal may be extended to a personal genome portal as a Korean personal genome browser. With additional variant annotation and related phenotypic information, the portal system will be able to provide personalized genomic profiles and many predictions about personal genomes. Overall, I presented the Korean genome analysis in the context of the Korean Genome Project (KGP), and its application in genomics, clinical practice, and population history.

References

1. Ezkurdia, I.; Juan, D.; Rodriguez, J. M.; Frankish, A.; Diekhans, M.; Harrow, J.; Vazquez, J.; Valencia, A.; Tress, M. L., Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum Mol Genet* **2014,** *23* (22), 5866-78.

2. Miga, K. H.; Koren, S.; Rhie, A.; Vollger, M. R.; Gershman, A.; Bzikadze, A.; Brooks, S.; Howe, E.; Porubsky, D.; Logsdon, G. A.; Schneider, V. A.; Potapova, T.; Wood, J.; Chow, W.; Armstrong, J.; Fredrickson, J.; Pak, E.; Tigyi, K.; Kremitzki, M.; Markovic, C.; Maduro, V.; Dutra, A.; Bouffard, G. G.; Chang, A. M.; Hansen, N. F.; Wilfert, A. B.; Thibaud-Nissen, F.; Schmitt, A. D.; Belton, J. M.; Selvaraj, S.; Dennis, M. Y.; Soto, D. C.; Sahasrabudhe, R.; Kaya, G.; Quick, J.; Loman, N. J.; Holmes, N.; Loose, M.; Surti, U.; Risques, R. A.; Graves Lindsay, T. A.; Fulton, R.; Hall, I.; Paten, B.; Howe, K.; Timp, W.; Young, A.; Mullikin, J. C.; Pevzner, P. A.; Gerton, J. L.; Sullivan, B. A.; Eichler, E. E.; Phillippy, A. M., Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **2020,** *585* (7823), 79-84.

3. Kim, S.; Cho, Y. S.; Kim, H. M.; Chung, O.; Kim, H.; Jho, S.; Seomun, H.; Kim, J.; Bang, W. Y.; Kim, C.; An, J.; Bae, C. H.; Bhak, Y.; Jeon, S.; Yoon, H.; Kim, Y.; Jun, J.; Lee, H.; Cho, S.; Uphyrkina, O.; Kostyria, A.; Goodrich, J.; Miquelle, D.; Roelke, M.; Lewis, J.; Yurchenko, A.; Bankevich, A.; Cho, J.; Lee, S.; Edwards, J. S.; Weber, J. A.; Cook, J.; Kim, S.; Lee, H.; Manica, A.; Lee, I.; O'Brien, S. J.; Bhak, J.; Yeo, J. H., Comparison of carnivore, omnivore, and herbivore mammalian genomes with a new leopard assembly. *Genome Biol* **2016,** *17* (1), 211.

4. Jeon, Y.; Park, S. G.; Lee, N.; Weber, J. A.; Kim, H. S.; Hwang, S. J.; Woo, S.; Kim, H. M.; Bhak, Y.; Jeon, S.; Lee, N.; Jo, Y.; Blazyte, A.; Ryu, T.; Cho, Y. S.; Kim, H.; Lee, J. H.; Yim, H. S.; Bhak, J.; Yum, S., The Draft Genome of an Octocoral, Dendronephthya gigantea. *Genome biology and evolution* **2019,** *11* (3), 949-953.

5. Kim, H. M.; Weber, J. A.; Lee, N.; Park, S. G.; Cho, Y. S.; Bhak, Y.; Lee, N.; Jeon, Y.; Jeon, S.; Luria, V.; Karger, A.; Kirschner, M. W.; Jo, Y. J.; Woo, S.; Shin, K.; Chung, O.; Ryu, J. C.; Yim, H. S.; Lee, J. H.; Edwards, J. S.; Manica, A.; Bhak, J.; Yum, S., The genome of the giant Nomura's jellyfish sheds light on the early evolution of active predation. *BMC Biol* **2019,** *17* (1), 28.

6. ElHefnawi, M.; Jeon, S.; Bhak, Y.; ElFiky, A.; Horaiz, A.; Jun, J.; Kim, H.; Bhak, J., Whole genome sequencing and bioinformatics analysis of two Egyptian genomes. *Gene* **2018,** *668*, 129-134.

7. Almal, S.; Jeon, S.; Agarwal, M.; Patel, S.; Patel, S.; Bhak, Y.; Jun, J.; Bhak, J.; Padh, H., Sequencing and analysis of the whole genome of Indian Gujarati male. *Genomics* **2019,** *111* (2), 196-204.

8. Seidualy, M.; Blazyte, A.; Jeon, S.; Bhak, Y.; Jeon, Y.; Kim, J.; Eriksson, A.; Bolser, D.; Yoon, C.; Manica, A.; Lee, S.; Bhak, J., Decoding a highly mixed Kazakh genome. *Hum Genet* **2020,** *139* (5), 557-568.

9. Consortium, T. G. P., A global reference for human genetic variation. *Nature* **2015,** *526* (7571), 68-74.

10. Karczewski, K. J.; Francioli, L. C.; Tiao, G.; Cummings, B. B.; Alfoldi, J.; Wang, Q.; Collins, R. L.; Laricchia, K. M.; Ganna, A.; Birnbaum, D. P.; Gauthier, L. D.; Brand, H.; Solomonson, M.; Watts, N. A.; Rhodes, D.; Singer-Berk, M.; England, E. M.; Seaby, E. G.; Kosmicki, J. A.; Walters, R. K.; Tashman, K.; Farjoun, Y.; Banks, E.; Poterba, T.; Wang, A.; Seed, C.; Whiffin, N.; Chong, J. X.; Samocha, K. E.; Pierce-Hoffman, E.; Zappala, Z.; O'Donnell-Luria, A. H.; Minikel, E. V.; Weisburd, B.; Lek, M.; Ware, J. S.; Vittal, C.; Armean, I. M.; Bergelson, L.; Cibulskis, K.; Connolly, K. M.; Covarrubias, M.; Donnelly, S.; Ferriera, S.; Gabriel, S.; Gentry, J.; Gupta, N.; Jeandet, T.; Kaplan, D.; Llanwarne, C.; Munshi, R.; Novod, S.; Petrillo, N.; Roazen, D.; Ruano-Rubio, V.; Saltzman, A.; Schleicher, M.; Soto, J.; Tibbetts, K.; Tolonen, C.; Wade, G.; Talkowski, M. E.; Genome Aggregation Database, C.; Neale, B. M.; Daly, M. J.; MacArthur, D. G., The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **2020,** *581* (7809), 434-443.

11. Siska, V.; Jones, E. R.; Jeon, S.; Bhak, Y.; Kim, H. M.; Cho, Y. S.; Kim, H.; Lee, K.; Veselovskaya, E.; Balueva, T.; Gallego-Llorente, M.; Hofreiter, M.; Bradley, D. G.; Eriksson, A.; Pinhasi, R.; Bhak, J.; Manica, A., Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago. *Sci Adv* **2017,** *3* (2), e1601877.

12. Barnett, R.; Westbury, M. V.; Sandoval-Velasco, M.; Vieira, F. G.; Jeon, S.; Zazula, G.; Martin, M. D.; Ho, S. Y. W.; Mather, N.; Gopalakrishnan, S.; Ramos-Madrigal, J.; de Manuel, M.; Zepeda-Mendoza, M. L.; Antunes, A.; Baez, A. C.; De Cahsan, B.; Larson, G.; O'Brien, S. J.; Eizirik, E.; Johnson, W. E.; Koepfli, K. P.; Wilting, A.; Fickel, J.; Dalen, L.; Lorenzen, E. D.; Marques-Bonet, T.; Hansen, A. J.; Zhang, G.; Bhak, J.; Yamaguchi, N.; Gilbert, M. T. P., Genomic Adaptations and Evolutionary History of the Extinct Scimitar-Toothed Cat, Homotherium latidens. *Current biology : CB* **2020,** *30* (24), 5018-5025 e5.

13. Ahn, S. M.; Kim, T. H.; Lee, S.; Kim, D.; Ghang, H.; Kim, D. S.; Kim, B. C.; Kim, S. Y.; Kim, W. Y.; Kim, C.; Park, D.; Lee, Y. S.; Kim, S.; Reja, R.; Jho, S.; Kim, C. G.; Cha, J. Y.; Kim, K. H.; Lee, B.; Bhak, J.; Kim, S. J., The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* **2009,** *19* (9), 1622-9.

14. Cho, Y. S.; Kim, H.; Kim, H. M.; Jho, S.; Jun, J.; Lee, Y. J.; Chae, K. S.; Kim, C. G.; Kim, S.; Eriksson, A.; Edwards, J. S.; Lee, S.; Kim, B. C.; Manica, A.; Oh, T. K.; Church, G. M.; Bhak, J., An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nat Commun* **2016,** *7*, 13637.

15. Kim, J.; Weber, J. A.; Jho, S.; Jang, J.; Jun, J.; Cho, Y. S.; Kim, H. M.; Kim, H.; Kim, Y.; Chung, O.; Kim, C. G.; Lee, H.; Kim, B. C.; Han, K.; Koh, I.; Chae, K. S.; Lee, S.; Edwards, J. S.; Bhak, J., KoVariome: Korean National Standard Reference Variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analyses. *Sci Rep* **2018,** *8* (1), 5677.

16. Lee, S.; Seo, J.; Park, J.; Nam, J. Y.; Choi, A.; Ignatius, J. S.; Bjornson, R. D.; Chae, J. H.; Jang, I. J.; Lee, S.; Park, W. Y.; Baek, D.; Choi, M., Korean Variant Archive (KOVA): a reference database of genetic variations in the Korean population. *Sci Rep* **2017,** *7* (1), 4287.

17. consortium, T. H.-P. A., Mapping Human Genetic Diversity in Asia. *Science* **2009,** *326* (5959), 1541-1545.

18. Databank, Population Total. **2018**.

19. Affairs, R. O. K. M. o. F., Total number of overseas Koreans. **2017**.

20. Seo, J. S.; Rhie, A.; Kim, J.; Lee, S.; Sohn, M. H.; Kim, C. U.; Hastie, A.; Cao, H.; Yun, J. Y.; Kim, J.; Kuk, J.; Park, G. H.; Kim, J.; Ryu, H.; Kim, J.; Roh, M.; Baek, J.; Hunkapiller, M. W.; Korlach, J.; Shin, J. Y.; Kim, C., De novo assembly and phasing of a Korean human genome. *Nature* **2016,** *538* (7624), 243-247.

21. Hong, D.; Park, S. S.; Ju, Y. S.; Kim, S.; Shin, J. Y.; Kim, S.; Yu, S. B.; Lee, W. C.; Lee, S.; Park, H.; Kim, J. I.; Seo, J. S., TIARA: a database for accurate analysis of multiple personal genomes based on cross-technology. *Nucleic Acids Res* **2011,** *39* (Database issue), D883-8.

22. Consortium, U. K.; Walter, K.; Min, J. L.; Huang, J.; Crooks, L.; Memari, Y.; McCarthy, S.; Perry, J. R.; Xu, C.; Futema, M.; Lawson, D.; Iotchkova, V.; Schiffels, S.; Hendricks, A. E.; Danecek, P.; Li, R.; Floyd, J.; Wain, L. V.; Barroso, I.; Humphries, S. E.; Hurles, M. E.; Zeggini, E.; Barrett, J. C.; Plagnol, V.; Richards, J. B.; Greenwood, C. M.; Timpson, N. J.; Durbin, R.; Soranzo, N., The UK10K project identifies rare variants in health and disease. *Nature* **2015,** *526* (7571), 82-90.

23. Genome of the Netherlands, C., Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **2014,** *46* (8), 818-25.

24. Sherman, R. M.; Forman, J.; Antonescu, V.; Puiu, D.; Daya, M.; Rafaels, N.; Boorgula, M. P.; Chavan, S.; Vergara, C.; Ortega, V. E.; Levin, A. M.; Eng, C.; Yazdanbakhsh, M.; Wilson, J. G.; Marrugo, J.; Lange, L. A.; Williams, L. K.; Watson, H.;

Ware, L. B.; Olopade, C. O.; Olopade, O.; Oliveira, R. R.; Ober, C.; Nicolae, D. L.; Meyers, D. A.; Mayorga, A.; Knight-Madden, J.; Hartert, T.; Hansel, N. N.; Foreman, M. G.; Ford, J. G.; Faruque, M. U.; Dunston, G. M.; Caraballo, L.; Burchard, E. G.; Bleecker, E. R.; Araujo, M. I.; Herrera-Paz, E. F.; Campbell, M.; Foster, C.; Taub, M. A.; Beaty, T. H.; Ruczinski, I.; Mathias, R. A.; Barnes, K. C.; Salzberg, S. L., Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet* **2019,** *51* (1), 30-35.

25. Gudbjartsson, D. F.; Helgason, H.; Gudjonsson, S. A.; Zink, F.; Oddson, A.; Gylfason, A.; Besenbacher, S.; Magnusson, G.; Halldorsson, B. V.; Hjartarson, E.; Sigurdsson, G. T.; Stacey, S. N.; Frigge, M. L.; Holm, H.; Saemundsdottir, J.; Helgadottir, H. T.; Johannsdottir, H.; Sigfusson, G.; Thorgeirsson, G.; Sverrisson, J. T.; Gretarsdottir, S.; Walters, G. B.; Rafnar, T.; Thjodleifsson, B.; Bjornsson, E. S.; Olafsson, S.; Thorarinsdottir, H.; Steingrimsdottir, T.; Gudmundsdottir, T. S.; Theodors, A.; Jonasson, J. G.; Sigurdsson, A.; Bjornsdottir, G.; Jonsson, J. J.; Thorarensen, O.; Ludvigsson, P.; Gudbjartsson, H.; Eyjolfsson, G. I.; Sigurdardottir, O.; Olafsson, I.; Arnar, D. O.; Magnusson, O. T.; Kong, A.; Masson, G.; Thorsteinsdottir, U.; Helgason, A.; Sulem, P.; Stefansson, K., Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **2015,** *47* (5), 435-44.

26. Maretty, L.; Jensen, J. M.; Petersen, B.; Sibbesen, J. A.; Liu, S.; Villesen, P.; Skov, L.; Belling, K.; Theil Have, C.; Izarzugaza, J. M. G.; Grosjean, M.; Bork-Jensen, J.; Grove, J.; Als, T. D.; Huang, S.; Chang, Y.; Xu, R.; Ye, W.; Rao, J.; Guo, X.; Sun, J.; Cao, H.; Ye, C.; van Beusekom, J.; Espeseth, T.; Flindt, E.; Friborg, R. M.; Halager, A. E.; Le Hellard, S.; Hultman, C. M.; Lescai, F.; Li, S.; Lund, O.; Longren, P.; Mailund, T.; Matey-Hernandez, M. L.; Mors, O.; Pedersen, C. N. S.; Sicheritz-Ponten, T.; Sullivan, P.; Syed, A.; Westergaard, D.; Yadav, R.; Li, N.; Xu, X.; Hansen, T.; Krogh, A.; Bolund, L.; Sorensen, T. I. A.; Pedersen, O.; Gupta, R.; Rasmussen, S.; Besenbacher, S.; Borglum, A. D.; Wang, J.; Eiberg, H.; Kristiansen, K.; Brunak, S.; Schierup, M. H., Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **2017,** *548* (7665), 87-91.

27. Nagasaki, M.; Yasuda, J.; Katsuoka, F.; Nariai, N.; Kojima, K.; Kawai, Y.; Yamaguchi-Kabata, Y.; Yokozawa, J.; Danjoh, I.; Saito, S.; Sato, Y.; Mimori, T.; Tsuda, K.; Saito, R.; Pan, X.; Nishikawa, S.; Ito, S.; Kuroki, Y.; Tanabe, O.; Fuse, N.; Kuriyama, S.; Kiyomoto, H.; Hozawa, A.; Minegishi, N.; Douglas Engel, J.; Kinoshita, K.; Kure, S.; Yaegashi, N.; To, M. J. R. P. P.; Yamamoto, M., Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun* **2015,** *6*, 8018.

28. Okada, Y.; Momozawa, Y.; Sakaue, S.; Kanai, M.; Ishigaki, K.; Akiyama, M.; Kishikawa, T.; Arai, Y.; Sasaki, T.; Kosaki, K.; Suematsu, M.; Matsuda, K.; Yamamoto,

K.; Kubo, M.; Hirose, N.; Kamatani, Y., Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nat Commun* **2018,** *9* (1), 1631.

29. Yoon, K.; Lee, S.; Han, T. S.; Moon, S. Y.; Yun, S. M.; Kong, S. H.; Jho, S.; Choe, J.; Yu, J.; Lee, H. J.; Park, J. H.; Kim, H. M.; Lee, S. Y.; Park, J.; Kim, W. H.; Bhak, J.; Yang, H. K.; Kim, S. J., Comprehensive genome- and transcriptome-wide analyses of mutations associated with microsatellite instability in Korean gastric cancers. *Genome Res* **2013,** *23* (7), 1109-17.

30. Martin, M., Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. *EMBnet.journal* **2011,** *17*.

31. Li, H.; Durbin, R., Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009,** *25* (14), 1754-60.

32. Ryan Poplin; Valentin Ruano-Rubio; Mark A. DePristo; Tim J. Fennell; Mauricio O. Carneiro; Geraldine A. Van der Auwera; David E. Kling; Laura D. Gauthier; Ami Levy-Moonshine; David Roazen; Khalid Shakir; Joel Thibault; Sheila Chandran; Chris Whelan; Monkol Lek; Stacey Gabriel; Mark J. Daly; Benjamin Neale; Daniel G. MacArthur; Banks, E., Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* **2017**.

33. McLaren, W.; Gil, L.; Hunt, S. E.; Riat, H. S.; Ritchie, G. R.; Thormann, A.; Flicek, P.; Cunningham, F., The Ensembl Variant Effect Predictor. *Genome Biol* **2016,** *17* (1), 122.

34. Moon, S.; Akey, J. M., A flexible method for estimating the fraction of fitness influencing mutations from large sequencing data sets. *Genome Res* **2016,** *26* (6), 834-43.

35. Abyzov, A.; Urban, A. E.; Snyder, M.; Gerstein, M., CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **2011,** *21* (6), 974-84.

36. Csardi, M. G., Package 'igraph'. *Last accessed* **2013,** *3* (09), 2013.

37. Zerbino, D. R.; Achuthan, P.; Akanni, W.; Amode, M. R.; Barrell, D.; Bhai, J.; Billis, K.; Cummins, C.; Gall, A.; Giron, C. G.; Gil, L.; Gordon, L.; Haggerty, L.; Haskell, E.; Hourlier, T.; Izuogu, O. G.; Janacek, S. H.; Juettemann, T.; To, J. K.; Laird, M. R.; Lavidas, I.; Liu, Z.; Loveland, J. E.; Maurel, T.; McLaren, W.; Moore, B.; Mudge, J.; Murphy, D. N.; Newman, V.; Nuhn, M.; Ogeh, D.; Ong, C. K.; Parker, A.; Patricio, M.; Riat, H. S.; Schuilenburg, H.; Sheppard, D.; Sparrow, H.; Taylor, K.; Thormann, A.; Vullo, A.; Walts, B.; Zadissa, A.; Frankish, A.; Hunt, S. E.; Kostadima, M.; Langridge, N.; Martin, F. J.; Muffato, M.; Perry, E.; Ruffier, M.; Staines, D. M.; Trevanion, S. J.; Aken, B. L.; Cunningham, F.; Yates, A.; Flicek, P., Ensembl 2018. *Nucleic Acids Res* **2018,** *46* (D1), D754-D761.

38. Sudmant, P. H.; Rausch, T.; Gardner, E. J.; Handsaker, R. E.; Abyzov, A.; Huddleston, J.; Zhang, Y.; Ye, K.; Jun, G.; Fritz, M. H.; Konkel, M. K.; Malhotra, A.;

Stutz, A. M.; Shi, X.; Casale, F. P.; Chen, J.; Hormozdiari, F.; Dayama, G.; Chen, K.; Malig, M.; Chaisson, M. J. P.; Walter, K.; Meiers, S.; Kashin, S.; Garrison, E.; Auton, A.; Lam, H. Y. K.; Mu, X. J.; Alkan, C.; Antaki, D.; Bae, T.; Cerveira, E.; Chines, P.; Chong, Z.; Clarke, L.; Dal, E.; Ding, L.; Emery, S.; Fan, X.; Gujral, M.; Kahveci, F.; Kidd, J. M.; Kong, Y.; Lameijer, E. W.; McCarthy, S.; Flicek, P.; Gibbs, R. A.; Marth, G.; Mason, C. E.; Menelaou, A.; Muzny, D. M.; Nelson, B. J.; Noor, A.; Parrish, N. F.; Pendleton, M.; Quitadamo, A.; Raeder, B.; Schadt, E. E.; Romanovitch, M.; Schlattl, A.; Sebra, R.; Shabalin, A. A.; Untergasser, A.; Walker, J. A.; Wang, M.; Yu, F.; Zhang, C.; Zhang, J.; Zheng-Bradley, X.; Zhou, W.; Zichner, T.; Sebat, J.; Batzer, M. A.; McCarroll, S. A.; Genomes Project, C.; Mills, R. E.; Gerstein, M. B.; Bashir, A.; Stegle, O.; Devine, S. E.; Lee, C.; Eichler, E. E.; Korbel, J. O., An integrated map of structural variation in 2,504 human genomes. *Nature* **2015,** *526* (7571), 75-81.

39. Boeva, V.; Popova, T.; Bleakley, K.; Chiche, P.; Cappo, J.; Schleiermacher, G.; Janoueix-Lerosey, I.; Delattre, O.; Barillot, E., Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **2012,** *28* (3), 423-5.

40. Gardner, E. J.; Lam, V. K.; Harris, D. N.; Chuang, N. T.; Scott, E. C.; Pittard, W. S.; Mills, R. E.; Genomes Project, C.; Devine, S. E., The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res* **2017,** *27* (11), 1916-1929.

41. Rishishwar, L.; Tellez Villa, C. E.; Jordan, I. K., Transposable element polymorphisms recapitulate human evolution. *Mob DNA* **2015,** *6*, 21.

42. Szolek, A.; Schubert, B.; Mohr, C.; Sturm, M.; Feldhahn, M.; Kohlbacher, O., OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **2014,** *30* (23), 3310-6.

43. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Genome Project Data Processing, S., The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009,** *25* (16), 2078-9.

44. Gonzalez-Galarza, F. F.; Takeshita, L. Y.; Santos, E. J.; Kempson, F.; Maia, M. H.; da Silva, A. L.; Teles e Silva, A. L.; Ghattaoraya, G. S.; Alfirevic, A.; Jones, A. R.; Middleton, D., Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res* **2015,** *43* (Database issue), D784-8.

45. Price, A. L.; Patterson, N. J.; Plenge, R. M.; Weinblatt, M. E.; Shadick, N. A.; Reich, D., Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **2006,** *38* (8), 904-9.

46. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M. A.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P. I.; Daly, M. J.; Sham, P. C., PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **2007,** *81* (3), 559-75.

47. Alexander, D. H.; Novembre, J.; Lange, K., Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **2009,** *19* (9), 1655-64.

48. Oven, M., PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Science International: Genetics Supplement Series* **2015,** *5*.

49. Weissensteiner, H.; Pacher, D.; Kloss-Brandstatter, A.; Forer, L.; Specht, G.; Bandelt, H. J.; Kronenberg, F.; Salas, A.; Schonherr, S., HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res* **2016,** *44* (W1), W58-63.

50. Jostins, L.; Xu, Y.; McCarthy, S.; Ayub, Q.; Durbin, R.; Barrett, J.; Tyler-Smith, C., YFitter: Maximum likelihood assignment of Y chromosome haplogroups from low-coverage sequence data. *arXiv preprint arXiv:1407.7988* **2014**.

51. Zhao, H.; Sun, Z.; Wang, J.; Huang, H.; Kocher, J. P.; Wang, L., CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **2014,** *30* (7), 1006-7.

52. MacArthur, J.; Bowler, E.; Cerezo, M.; Gil, L.; Hall, P.; Hastings, E.; Junkins, H.; McMahon, A.; Milano, A.; Morales, J.; Pendlington, Z. M.; Welter, D.; Burdett, T.; Hindorff, L.; Flicek, P.; Cunningham, F.; Parkinson, H., The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **2017,** *45* (D1), D896-D901.

53. McCarthy, S.; Das, S.; Kretzschmar, W.; Delaneau, O.; Wood, A. R.; Teumer, A.; Kang, H. M.; Fuchsberger, C.; Danecek, P.; Sharp, K.; Luo, Y.; Sidore, C.; Kwong, A.; Timpson, N.; Koskinen, S.; Vrieze, S.; Scott, L. J.; Zhang, H.; Mahajan, A.; Veldink, J.; Peters, U.; Pato, C.; van Duijn, C. M.; Gillies, C. E.; Gandin, I.; Mezzavilla, M.; Gilly, A.; Cocca, M.; Traglia, M.; Angius, A.; Barrett, J. C.; Boomsma, D.; Branham, K.; Breen, G.; Brummett, C. M.; Busonero, F.; Campbell, H.; Chan, A.; Chen, S.; Chew, E.; Collins, F. S.; Corbin, L. J.; Smith, G. D.; Dedoussis, G.; Dorr, M.; Farmaki, A. E.; Ferrucci, L.; Forer, L.; Fraser, R. M.; Gabriel, S.; Levy, S.; Groop, L.; Harrison, T.; Hattersley, A.; Holmen, O. L.; Hveem, K.; Kretzler, M.; Lee, J. C.; McGue, M.; Meitinger, T.; Melzer, D.; Min, J. L.; Mohlke, K. L.; Vincent, J. B.; Nauck, M.; Nickerson, D.; Palotie, A.; Pato, M.; Pirastu, N.; McInnis, M.; Richards, J. B.; Sala, C.; Salomaa, V.; Schlessinger, D.; Schoenherr, S.; Slagboom, P. E.; Small, K.; Spector, T.; Stambolian, D.; Tuke, M.; Tuomilehto, J.; Van den Berg, L. H.; Van Rheenen, W.; Volker, U.; Wijmenga, C.; Toniolo, D.; Zeggini, E.; Gasparini, P.; Sampson, M. G.; Wilson, J. F.; Frayling, T.; de Bakker, P. I.; Swertz, M. A.; McCarroll, S.; Kooperberg, C.; Dekker, A.; Altshuler, D.; Willer, C.; Iacono, W.; Ripatti, S.; Soranzo, N.; Walter, K.; Swaroop, A.; Cucca, F.; Anderson, C. A.; Myers, R. M.; Boehnke, M.; McCarthy, M. I.; Durbin, R.; Haplotype Reference, C., A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **2016,** *48* (10), 1279-83.

54. Das, S.; Forer, L.; Schonherr, S.; Sidore, C.; Locke, A. E.; Kwong, A.; Vrieze, S. I.; Chew, E. Y.; Levy, S.; McGue, M.; Schlessinger, D.; Stambolian, D.; Loh, P. R.; Iacono, W. G.; Swaroop, A.; Scott, L. J.; Cucca, F.; Kronenberg, F.; Boehnke, M.; Abecasis, G. R.; Fuchsberger, C., Next-generation genotype imputation service and methods. *Nat Genet* **2016,** *48* (10), 1284-1287.

55. Sherry, S. T.; Ward, M. H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E. M.; Sirotkin, K., dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **2001,** *29* (1), 308-11.

56. Clark, T. G.; Andrew, T.; Cooper, G. M.; Margulies, E. H.; Mullikin, J. C.; Balding, D. J., Functional constraint and small insertions and deletions in the ENCODE regions of the human genome. *Genome Biol* **2007,** *8* (9), R180.

57. Telenti, A.; Pierce, L. C.; Biggs, W. H.; di Iulio, J.; Wong, E. H.; Fabani, M. M.; Kirkness, E. F.; Moustafa, A.; Shah, N.; Xie, C.; Brewerton, S. C.; Bulsara, N.; Garner, C.; Metzker, G.; Sandoval, E.; Perkins, B. A.; Och, F. J.; Turpaz, Y.; Venter, J. C., Deep sequencing of 10,000 human genomes. *Proc Natl Acad Sci U S A* **2016,** *113* (42), 11901-11906.

58. Adzhubei, I.; Jordan, D. M.; Sunyaev, S. R., Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **2013,** *Chapter 7*, Unit7 20.

59. Sim, N. L.; Kumar, P.; Hu, J.; Henikoff, S.; Schneider, G.; Ng, P. C., SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* **2012,** *40* (Web Server issue), W452-7.

60. Jin, H. J.; Kwak, K. D.; Hammer, M. F.; Nakahori, Y.; Shinka, T.; Lee, J. W.; Jin, F.; Jia, X.; Tyler-Smith, C.; Kim, W., Y-chromosomal DNA haplogroups and their implications for the dual origins of the Koreans. *Hum Genet* **2003,** *114* (1), 27-35.

61. Jin, H. J.; Tyler-Smith, C.; Kim, W., The peopling of Korea revealed by analyses of mitochondrial DNA and Y-chromosomal markers. *PLoS One* **2009,** *4* (1), e4210.

62. Tanaka, M.; Cabrera, V. M.; Gonzalez, A. M.; Larruga, J. M.; Takeyasu, T.; Fuku, N.; Guo, L. J.; Hirose, R.; Fujita, Y.; Kurata, M.; Shinoda, K.; Umetsu, K.; Yamada, Y.; Oshida, Y.; Sato, Y.; Hattori, N.; Mizuno, Y.; Arai, Y.; Hirose, N.; Ohta, S.; Ogawa, O.; Tanaka, Y.; Kawamori, R.; Shamoto-Nagai, M.; Maruyama, W.; Shimokata, H.; Suzuki, R.; Shimodaira, H., Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res* **2004,** *14* (10A), 1832-50.

63. Wang, Y.; Lu, D.; Chung, Y. J.; Xu, S., Genetic structure, divergence and admixture of Han Chinese, Japanese and Korean populations. *Hereditas* **2018,** *155*, 19.

64. Cusi, D.; Barlassina, C.; Azzani, T.; Casari, G.; Citterio, L.; Devoto, M.; Glorioso, N.; Lanzani, C.; Manunta, P.; Righetti, M.; Rivera, R.; Stella, P.; Troffa, C.; Zagato, L.; Bianchi, G., Polymorphisms of alpha-adducin and salt sensitivity in patients with essential hypertension. *Lancet* **1997,** *349* (9062), 1353-7.

65. Psaty, B. M.; Smith, N. L.; Heckbert, S. R.; Vos, H. L.; Lemaitre, R. N.; Reiner, A. P.; Siscovick, D. S.; Bis, J.; Lumley, T.; Longstreth, W. T., Jr.; Rosendaal, F. R., Diuretic therapy, the alpha-adducin gene variant, and the risk of myocardial infarction or stroke in persons with treated hypertension. *JAMA* **2002,** *287* (13), 1680-9.

66. Genomes Project, C.; Abecasis, G. R.; Altshuler, D.; Auton, A.; Brooks, L. D.; Durbin, R. M.; Gibbs, R. A.; Hurles, M. E.; McVean, G. A., A map of human genome variation from population-scale sequencing. *Nature* **2010,** *467* (7319), 1061-73.

67. Bycroft, C.; Freeman, C.; Petkova, D.; Band, G.; Elliott, L. T.; Sharp, K.; Motyer, A.; Vukcevic, D.; Delaneau, O.; O'Connell, J.; Cortes, A.; Welsh, S.; Young, A.; Effingham, M.; McVean, G.; Leslie, S.; Allen, N.; Donnelly, P.; Marchini, J., The UK Biobank resource with deep phenotyping and genomic data. *Nature* **2018,** *562* (7726), 203-209.

68. Kang, T. W.; Kim, H. J.; Ju, H.; Kim, J. H.; Jeon, Y. J.; Lee, H. C.; Kim, K. K.; Kim, J. W.; Lee, S.; Kim, J. Y.; Kim, S. Y.; Kim, Y. S., Genome-wide association of serum bilirubin levels in Korean population. *Hum Mol Genet* **2010,** *19* (18), 3672-8.

69. Kim, Y. J.; Go, M. J.; Hu, C.; Hong, C. B.; Kim, Y. K.; Lee, J. Y.; Hwang, J. Y.; Oh, J. H.; Kim, D. J.; Kim, N. H.; Kim, S.; Hong, E. J.; Kim, J. H.; Min, H.; Kim, Y.; Zhang, R.; Jia, W.; Okada, Y.; Takahashi, A.; Kubo, M.; Tanaka, T.; Kamatani, N.; Matsuda, K.; consortium, M.; Park, T.; Oh, B.; Kimm, K.; Kang, D.; Shin, C.; Cho, N. H.; Kim, H. L.; Han, B. G.; Lee, J. Y.; Cho, Y. S., Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits. *Nat Genet* **2011,** *43* (10), 990-5.

70. Consortium, H. P.-A. S.; Abdulla, M. A.; Ahmed, I.; Assawamakin, A.; Bhak, J.; Brahmachari, S. K.; Calacal, G. C.; Chaurasia, A.; Chen, C. H.; Chen, J.; Chen, Y. T.; Chu, J.; Cutiongco-de la Paz, E. M.; De Ungria, M. C.; Delfin, F. C.; Edo, J.; Fuchareon, S.; Ghang, H.; Gojobori, T.; Han, J.; Ho, S. F.; Hoh, B. P.; Huang, W.; Inoko, H.; Jha, P.; Jinam, T. A.; Jin, L.; Jung, J.; Kangwanpong, D.; Kampuansai, J.; Kennedy, G. C.; Khurana, P.; Kim, H. L.; Kim, K.; Kim, S.; Kim, W. Y.; Kimm, K.; Kimura, R.; Koike, T.; Kulawonganunchai, S.; Kumar, V.; Lai, P. S.; Lee, J. Y.; Lee, S.; Liu, E. T.; Majumder, P. P.; Mandapati, K. K.; Marzuki, S.; Mitchell, W.; Mukerji, M.; Naritomi, K.; Ngamphiw, C.; Niikawa, N.; Nishida, N.; Oh, B.; Oh, S.; Ohashi, J.; Oka, A.; Ong, R.; Padilla, C. D.; Palittapongarnpim, P.; Perdigon, H. B.; Phipps, M. E.; Png, E.; Sakaki, Y.; Salvador, J. M.; Sandraling, Y.; Scaria, V.; Seielstad, M.; Sidek, M. R.; Sinha, A.; Srikummool, M.; Sudoyo, H.; Sugano, S.; Suryadi, H.; Suzuki, Y.; Tabbada, K. A.; Tan, A.; Tokunaga, K.; Tongsima, S.; Villamor, L. P.; Wang, E.; Wang, Y.; Wang, H.; Wu, J. Y.; Xiao, H.; Xu, S.; Yang, J. O.; Shugart, Y. Y.; Yoo, H. S.; Yuan, W.; Zhao, G.; Zilfalil, B. A.; Indian Genome Variation, C., Mapping human genetic diversity in Asia. *Science* **2009,** *326* (5959), 1541-5.

71. Liu, X.; Lu, D.; Saw, W. Y.; Shaw, P. J.; Wangkumhang, P.; Ngamphiw, C.; Fucharoen, S.; Lert-Itthiporn, W.; Chin-Inmanu, K.; Chau, T. N.; Anders, K.; Kasturiratne, A.; de Silva, H. J.; Katsuya, T.; Kimura, R.; Nabika, T.; Ohkubo, T.; Tabara, Y.; Takeuchi, F.; Yamamoto, K.; Yokota, M.; Mamatyusupu, D.; Yang, W.; Chung, Y. J.; Jin, L.; Hoh, B. P.; Wickremasinghe, A. R.; Ong, R. H.; Khor, C. C.; Dunstan, S. J.; Simmons, C.; Tongsima, S.; Suriyaphol, P.; Kato, N.; Xu, S.; Teo, Y. Y., Characterising private and shared signatures of positive selection in 37 Asian populations. *European journal of human genetics : EJHG* **2017,** *25* (4), 499-508.

72. Kim, S.-H.; Kim, K.-C.; Shin, D.-J.; Jin, H.-J.; Kwak, K.-D.; Han, M.-S.; Song, J.-M.; Kim, W.; Kim, W., High frequencies of Y-chromosome haplogroup O2b-SRY465 lineages in Korea: a genetic perspective on the peopling of Korea. *Investigative Genetics* **2011,** *2*, 10-10.

73. Jin, H.-J.; Tyler-Smith, C.; Kim, W., The peopling of Korea revealed by analyses of mitochondrial DNA and Y-chromosomal markers. *PloS one* **2009,** *4* (1), e4210-e4210.

74. Takeuchi, F.; Katsuya, T.; Kimura, R.; Nabika, T.; Isomura, M.; Ohkubo, T.; Tabara, Y.; Yamamoto, K.; Yokota, M.; Liu, X.; Saw, W. Y.; Mamatyusupu, D.; Yang, W.; Xu, S.; Japanese Genome Variation, C.; Teo, Y. Y.; Kato, N., The fine-scale genetic structure and evolution of the Japanese population. *PLoS One* **2017,** *12* (11), e0185487.

75. Norton, C. J., The current state of korean paleoanthropology. *Journal of human evolution* **2000,** *38* (6), 803-25.

76. Park, Y. C., Chronology of palaeolithic sites and its cultural transition in Korea. *J Korean Archaeol. Soc.* **1992,** *28*, 5-130.

77. Bae, C. J.; Bae, K., The nature of the Early to Late Paleolithic transition in Korea: Current perspectives. *Quaternary International* **2012,** *281*, 26-35.

78. Yang, M. A.; Gao, X.; Theunert, C.; Tong, H.; Aximu-Petri, A.; Nickel, B.; Slatkin, M.; Meyer, M.; Paabo, S.; Kelso, J.; Fu, Q., 40,000-Year-Old Individual from Asia Provides Insight into Early Population Structure in Eurasia. *Current biology : CB* **2017,** *27* (20), 3202-3208 e9.

79. Lipson, M.; Cheronet, O.; Mallick, S.; Rohland, N.; Oxenham, M.; Pietrusewsky, M.; Pryce, T. O.; Willis, A.; Matsumura, H.; Buckley, H.; Domett, K.; Nguyen, G. H.; Trinh, H. H.; Kyaw, A. A.; Win, T. T.; Pradier, B.; Broomandkhoshbacht, N.; Candilio, F.; Changmai, P.; Fernandes, D.; Ferry, M.; Gamarra, B.; Harney, E.; Kampuansai, J.; Kutanan, W.; Michel, M.; Novak, M.; Oppenheimer, J.; Sirak, K.; Stewardson, K.; Zhang, Z.; Flegontov, P.; Pinhasi, R.; Reich, D., Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science* **2018,** *361* (6397), 92-95.

80. Lazaridis, I.; Patterson, N.; Mittnik, A.; Renaud, G.; Mallick, S.; Kirsanow, K.; Sudmant, P. H.; Schraiber, J. G.; Castellano, S.; Lipson, M.; Berger, B.; Economou, C.;

Bollongino, R.; Fu, Q.; Bos, K. I.; Nordenfelt, S.; Li, H.; de Filippo, C.; Prufer, K.; Sawyer, S.; Posth, C.; Haak, W.; Hallgren, F.; Fornander, E.; Rohland, N.; Delsate, D.; Francken, M.; Guinet, J. M.; Wahl, J.; Ayodo, G.; Babiker, H. A.; Bailliet, G.; Balanovska, E.; Balanovsky, O.; Barrantes, R.; Bedoya, G.; Ben-Ami, H.; Bene, J.; Berrada, F.; Bravi, C. M.; Brisighelli, F.; Busby, G. B.; Cali, F.; Churnosov, M.; Cole, D. E.; Corach, D.; Damba, L.; van Driem, G.; Dryomov, S.; Dugoujon, J. M.; Fedorova, S. A.; Gallego Romero, I.; Gubina, M.; Hammer, M.; Henn, B. M.; Hervig, T.; Hodoglugil, U.; Jha, A. R.; Karachanak-Yankova, S.; Khusainova, R.; Khusnutdinova, E.; Kittles, R.; Kivisild, T.; Klitz, W.; Kucinskas, V.; Kushniarevich, A.; Laredj, L.; Litvinov, S.; Loukidis, T.; Mahley, R. W.; Melegh, B.; Metspalu, E.; Molina, J.; Mountain, J.; Nakkalajarvi, K.; Nesheva, D.; Nyambo, T.; Osipova, L.; Parik, J.; Platonov, F.; Posukh, O.; Romano, V.; Rothhammer, F.; Rudan, I.; Ruizbakiev, R.; Sahakyan, H.; Sajantila, A.; Salas, A.; Starikovskaya, E. B.; Tarekegn, A.; Toncheva, D.; Turdikulova, S.; Uktveryte, I.; Utevska, O.; Vasquez, R.; Villena, M.; Voevoda, M.; Winkler, C. A.; Yepiskoposyan, L.; Zalloua, P.; Zemunik, T.; Cooper, A.; Capelli, C.; Thomas, M. G.; Ruiz-Linares, A.; Tishkoff, S. A.; Singh, L.; Thangaraj, K.; Villems, R.; Comas, D.; Sukernik, R.; Metspalu, M.; Meyer, M.; Eichler, E. E.; Burger, J.; Slatkin, M.; Paabo, S.; Kelso, J.; Reich, D.; Krause, J., Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **2014,** *513* (7518), 409-13.

81. Patel, R. K.; Jain, M., NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* **2012,** *7* (2), e30619.

82. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; DePristo, M. A., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **2010,** *20* (9), 1297-303.

83. Kloss-Brandstatter, A.; Pacher, D.; Schonherr, S.; Weissensteiner, H.; Binna, R.; Specht, G.; Kronenberg, F., HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum Mutat* **2011,** *32* (1), 25-32.

84. Bonatto, S. L.; Salzano, F. M., Diversity and age of the four major mtDNA haplogroups, and their implications for the peopling of the New World. *Am J Hum Genet* **1997,** *61* (6), 1413-23.

85. Lawson, D. J.; Hellenthal, G.; Myers, S.; Falush, D., Inference of population structure using dense haplotype data. *PLoS genetics* **2012,** *8* (1), e1002453.

86. Patterson, N.; Price, A. L.; Reich, D., Population structure and eigenanalysis. *PLoS genetics* **2006,** *2* (12), e190.

87. Loh, P. R.; Lipson, M.; Patterson, N.; Moorjani, P.; Pickrell, J. K.; Reich, D.; Berger, B., Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **2013,** *193* (4), 1233-54.

88. Patterson, N.; Moorjani, P.; Luo, Y.; Mallick, S.; Rohland, N.; Zhan, Y.; Genschoreck, T.; Webster, T.; Reich, D., Ancient admixture in human history. *Genetics* **2012,** *192* (3), 1065-93.

89. McColl, H.; Racimo, F.; Vinner, L.; Demeter, F.; Gakuhari, T.; Moreno-Mayar, J. V.; van Driem, G.; Gram Wilken, U.; Seguin-Orlando, A.; de la Fuente Castro, C.; Wasef, S.; Shoocongdej, R.; Souksavatdy, V.; Sayavongkhamdy, T.; Saidin, M. M.; Allentoft, M. E.; Sato, T.; Malaspinas, A. S.; Aghakhanian, F. A.; Korneliussen, T.; Prohaska, A.; Margaryan, A.; de Barros Damgaard, P.; Kaewsutthi, S.; Lertrit, P.; Nguyen, T. M. H.; Hung, H. C.; Minh Tran, T.; Nghia Truong, H.; Nguyen, G. H.; Shahidan, S.; Wiradnyana, K.; Matsumae, H.; Shigehara, N.; Yoneda, M.; Ishida, H.; Masuyama, T.; Yamada, Y.; Tajima, A.; Shibata, H.; Toyoda, A.; Hanihara, T.; Nakagome, S.; Deviese, T.; Bacon, A. M.; Duringer, P.; Ponche, J. L.; Shackelford, L.; Patole-Edoumba, E.; Nguyen, A. T.; Bellina-Pryce, B.; Galipaud, J. C.; Kinaston, R.; Buckley, H.; Pottier, C.; Rasmussen, S.; Higham, T.; Foley, R. A.; Lahr, M. M.; Orlando, L.; Sikora, M.; Phipps, M. E.; Oota, H.; Higham, C.; Lambert, D. M.; Willerslev, E., The prehistoric peopling of Southeast Asia. *Science* **2018,** *361* (6397), 88-92.

90. Kim, Y. J.; Jin, H. J., Dissecting the genetic structure of Korean population using genome-wide SNP arrays. *Genes & Genomics* **2013,** *35* (3), 355-363.

91. Haak, W.; Lazaridis, I.; Patterson, N.; Rohland, N.; Mallick, S.; Llamas, B.; Brandt, G.; Nordenfelt, S.; Harney, E.; Stewardson, K.; Fu, Q.; Mittnik, A.; Banffy, E.; Economou, C.; Francken, M.; Friederich, S.; Pena, R. G.; Hallgren, F.; Khartanovich, V.; Khokhlov, A.; Kunst, M.; Kuznetsov, P.; Meller, H.; Mochalov, O.; Moiseyev, V.; Nicklisch, N.; Pichler, S. L.; Risch, R.; Rojo Guerra, M. A.; Roth, C.; Szecsenyi-Nagy, A.; Wahl, J.; Meyer, M.; Krause, J.; Brown, D.; Anthony, D.; Cooper, A.; Alt, K. W.; Reich, D., Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **2015,** *522* (7555), 207-11.

92. de Barros Damgaard, P.; Martiniano, R.; Kamm, J.; Moreno-Mayar, J. V.; Kroonen, G.; Peyrot, M.; Barjamovic, G.; Rasmussen, S.; Zacho, C.; Baimukhanov, N.; Zaibert, V.; Merz, V.; Biddanda, A.; Merz, I.; Loman, V.; Evdokimov, V.; Usmanova, E.; Hemphill, B.; Seguin-Orlando, A.; Yediay, F. E.; Ullah, I.; Sjogren, K. G.; Iversen, K. H.; Choin, J.; de la Fuente, C.; Ilardo, M.; Schroeder, H.; Moiseyev, V.; Gromov, A.; Polyakov, A.; Omura, S.; Senyurt, S. Y.; Ahmad, H.; McKenzie, C.; Margaryan, A.; Hameed, A.; Samad, A.; Gul, N.; Khokhar, M. H.; Goriunova, O. I.; Bazaliiskii, V. I.; Novembre, J.; Weber, A.

W.; Orlando, L.; Allentoft, M. E.; Nielsen, R.; Kristiansen, K.; Sikora, M.; Outram, A. K.; Durbin, R.; Willerslev, E., The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* **2018,** *360* (6396).

93. Allentoft, M. E.; Sikora, M.; Sjogren, K. G.; Rasmussen, S.; Rasmussen, M.; Stenderup, J.; Damgaard, P. B.; Schroeder, H.; Ahlstrom, T.; Vinner, L.; Malaspinas, A. S.; Margaryan, A.; Higham, T.; Chivall, D.; Lynnerup, N.; Harvig, L.; Baron, J.; Della Casa, P.; Dabrowski, P.; Duffy, P. R.; Ebel, A. V.; Epimakhov, A.; Frei, K.; Furmanek, M.; Gralak, T.; Gromov, A.; Gronkiewicz, S.; Grupe, G.; Hajdu, T.; Jarysz, R.; Khartanovich, V.; Khokhlov, A.; Kiss, V.; Kolar, J.; Kriiska, A.; Lasak, I.; Longhi, C.; McGlynn, G.; Merkevicius, A.; Merkyte, I.; Metspalu, M.; Mkrtchyan, R.; Moiseyev, V.; Paja, L.; Palfi, G.; Pokutta, D.; Pospieszny, L.; Price, T. D.; Saag, L.; Sablin, M.; Shishlina, N.; Smrcka, V.; Soenov, V. I.; Szeverenyi, V.; Toth, G.; Trifanova, S. V.; Varul, L.; Vicze, M.; Yepiskoposyan, L.; Zhitenev, V.; Orlando, L.; Sicheritz-Ponten, T.; Brunak, S.; Nielsen, R.; Kristiansen, K.; Willerslev, E., Population genomics of Bronze Age Eurasia. *Nature* **2015,** *522* (7555), 167-72.

94. Damgaard, P. B.; Marchi, N.; Rasmussen, S.; Peyrot, M.; Renaud, G.; Korneliussen, T.; Moreno-Mayar, J. V.; Pedersen, M. W.; Goldberg, A.; Usmanova, E.; Baimukhanov, N.; Loman, V.; Hedeager, L.; Pedersen, A. G.; Nielsen, K.; Afanasiev, G.; Akmatov, K.; Aldashev, A.; Alpaslan, A.; Baimbetov, G.; Bazaliiskii, V. I.; Beisenov, A.; Boldbaatar, B.; Boldgiv, B.; Dorzhu, C.; Ellingvag, S.; Erdenebaatar, D.; Dajani, R.; Dmitriev, E.; Evdokimov, V.; Frei, K. M.; Gromov, A.; Goryachev, A.; Hakonarson, H.; Hegay, T.; Khachatryan, Z.; Khaskhanov, R.; Kitov, E.; Kolbina, A.; Kubatbek, T.; Kukushkin, A.; Kukushkin, I.; Lau, N.; Margaryan, A.; Merkyte, I.; Mertz, I. V.; Mertz, V. K.; Mijiddorj, E.; Moiyesev, V.; Mukhtarova, G.; Nurmukhanbetov, B.; Orozbekova, Z.; Panyushkina, I.; Pieta, K.; Smrcka, V.; Shevnina, I.; Logvin, A.; Sjogren, K. G.; Stolcova, T.; Taravella, A. M.; Tashbaeva, K.; Tkachev, A.; Tulegenov, T.; Voyakin, D.; Yepiskoposyan, L.; Undrakhbold, S.; Varfolomeev, V.; Weber, A.; Wilson Sayres, M. A.; Kradin, N.; Allentoft, M. E.; Orlando, L.; Nielsen, R.; Sikora, M.; Heyer, E.; Kristiansen, K.; Willerslev, E., 137 ancient human genomes from across the Eurasian steppes. *Nature* **2018,** *557* (7705), 369-374.

95. Chiaroni, J.; Underhill, P. A.; Cavalli-Sforza, L. L., Y chromosome diversity, human expansion, drift, and cultural evolution. *Proc Natl Acad Sci U S A* **2009,** *106* (48), 20174-9.

96. Karmin, M.; Saag, L.; Vicente, M.; Wilson Sayres, M. A.; Jarve, M.; Talas, U. G.; Rootsi, S.; Ilumae, A. M.; Magi, R.; Mitt, M.; Pagani, L.; Puurand, T.; Faltyskova, Z.; Clemente, F.; Cardona, A.; Metspalu, E.; Sahakyan, H.; Yunusbayev, B.; Hudjashov, G.; DeGiorgio, M.; Loogvali, E. L.; Eichstaedt, C.; Eelmets, M.; Chaubey, G.; Tambets, K.;

Litvinov, S.; Mormina, M.; Xue, Y.; Ayub, Q.; Zoraqi, G.; Korneliussen, T. S.; Akhatova, F.; Lachance, J.; Tishkoff, S.; Momynaliev, K.; Ricaut, F. X.; Kusuma, P.; Razafindrazaka, H.; Pierron, D.; Cox, M. P.; Sultana, G. N.; Willerslev, R.; Muller, C.; Westaway, M.; Lambert, D.; Skaro, V.; Kovacevic, L.; Turdikulova, S.; Dalimova, D.; Khusainova, R.; Trofimova, N.; Akhmetova, V.; Khidiyatova, I.; Lichman, D. V.; Isakova, J.; Pocheshkhova, E.; Sabitov, Z.; Barashkov, N. A.; Nymadawa, P.; Mihailov, E.; Seng, J. W.; Evseeva, I.; Migliano, A. B.; Abdullah, S.; Andriadze, G.; Primorac, D.; Atramentova, L.; Utevska, O.; Yepiskoposyan, L.; Marjanovic, D.; Kushniarevich, A.; Behar, D. M.; Gilissen, C.; Vissers, L.; Veltman, J. A.; Balanovska, E.; Derenko, M.; Malyarchuk, B.; Metspalu, A.; Fedorova, S.; Eriksson, A.; Manica, A.; Mendez, F. L.; Karafet, T. M.; Veeramah, K. R.; Bradman, N.; Hammer, M. F.; Osipova, L. P.; Balanovsky, O.; Khusnutdinova, E. K.; Johnsen, K.; Remm, M.; Thomas, M. G.; Tyler-Smith, C.; Underhill, P. A.; Willerslev, E.; Nielsen, R.; Metspalu, M.; Villems, R.; Kivisild, T., A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res* **2015,** *25* (4), 459-66.

97. Soares, P.; Ermini, L.; Thomson, N.; Mormina, M.; Rito, T.; Rohl, A.; Salas, A.; Oppenheimer, S.; Macaulay, V.; Richards, M. B., Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet* **2009,** *84* (6), 740-59.

98. Sayers, E. W.; Beck, J.; Bolton, E. E.; Bourexis, D.; Brister, J. R.; Canese, K.; Comeau, D. C.; Funk, K.; Kim, S.; Klimke, W.; Marchler-Bauer, A.; Landrum, M.; Lathrop, S.; Lu, Z.; Madden, T. L.; O'Leary, N.; Phan, L.; Rangwala, S. H.; Schneider, V. A.; Skripchenko, Y.; Wang, J.; Ye, J.; Trawick, B. W.; Pruitt, K. D.; Sherry, S. T., Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **2021,** *49* (D1), D10-D17.

99. Kent, W. J.; Sugnet, C. W.; Furey, T. S.; Roskin, K. M.; Pringle, T. H.; Zahler, A. M.; Haussler, D., The human genome browser at UCSC. *Genome Res* **2002,** *12* (6), 996-1006.

100. Yates, A. D.; Achuthan, P.; Akanni, W.; Allen, J.; Allen, J.; Alvarez-Jarreta, J.; Amode, M. R.; Armean, I. M.; Azov, A. G.; Bennett, R.; Bhai, J.; Billis, K.; Boddu, S.; Marugan, J. C.; Cummins, C.; Davidson, C.; Dodiya, K.; Fatima, R.; Gall, A.; Giron, C. G.; Gil, L.; Grego, T.; Haggerty, L.; Haskell, E.; Hourlier, T.; Izuogu, O. G.; Janacek, S. H.; Juettemann, T.; Kay, M.; Lavidas, I.; Le, T.; Lemos, D.; Martinez, J. G.; Maurel, T.; McDowall, M.; McMahon, A.; Mohanan, S.; Moore, B.; Nuhn, M.; Oheh, D. N.; Parker, A.; Parton, A.; Patricio, M.; Sakthivel, M. P.; Abdul Salam, A. I.; Schmitt, B. M.; Schuilenburg, H.; Sheppard, D.; Sycheva, M.; Szuba, M.; Taylor, K.; Thormann, A.; Threadgold, G.; Vullo, A.; Walts, B.; Winterbottom, A.; Zadissa, A.; Chakiachvili, M.; Flint, B.; Frankish, A.; Hunt, S. E.; G, I. I.; Kostadima, M.; Langridge, N.; Loveland, J. E.; Martin, F. J.; Morales, J.; Mudge, J. M.; Muffato, M.; Perry, E.; Ruffier, M.; Trevanion, S.

J.; Cunningham, F.; Howe, K. L.; Zerbino, D. R.; Flicek, P., Ensembl 2020. *Nucleic Acids Res* **2020,** *48* (D1), D682-D688.

101. Karczewski, K. J.; Weisburd, B.; Thomas, B.; Solomonson, M.; Ruderfer, D. M.; Kavanagh, D.; Hamamsy, T.; Lek, M.; Samocha, K. E.; Cummings, B. B.; Birnbaum, D.; The Exome Aggregation, C.; Daly, M. J.; MacArthur, D. G., The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res* **2017,** *45* (D1), D840-D845.

102. Zhang, C.; Gao, Y.; Ning, Z.; Lu, Y.; Zhang, X.; Liu, J.; Xie, B.; Xue, Z.; Wang, X.; Yuan, K.; Ge, X.; Pan, Y.; Liu, C.; Tian, L.; Wang, Y.; Lu, D.; Hoh, B. P.; Xu, S., PGG.SNV: understanding the evolutionary and medical implications of human single nucleotide variations in diverse populations. *Genome Biol* **2019,** *20* (1), 215.

103. Jung, K. S.; Hong, K. W.; Jo, H. Y.; Choi, J.; Ban, H. J.; Cho, S. B.; Chung, M., KRGDB: the large-scale variant database of 1722 Koreans based on whole genome sequencing. *Database (Oxford)* **2020,** *2020*.

Acknowledgments

Firstly, I appreciate all participants of Genome Korea, which is the representative Korean Genome Project (KGP) and Ulsan citizens supporting the Genome Korea in Ulsan project.

I sincerely thank Professor Jong Bhak, who supervised me and provided a lot of opportunities of doing genomics and bioinformatics research. He supervised me to think about what science is and what scientist and bioinformatists need to do. He has always treated me as an equal colleague rather than his student. This makes me able to investigate any topics freely and scientifically without any restrictions or pressure.

I thank Professor Semin Lee, who also guided and encourage me a lot. I also thank other committee members, Professor Seung Woo Cho, Professor Dougu Nam, and Dr. Jeongeun Kim, for their critical and meaningful comments and suggestions.

I also thank Dr. Hak-Min Kim, who motivate me to learn bio programming and develop many scripts by myself. Thanks to this, I can achieve the ability to create efficient bioinformatics pipelines and analyses various bio-omics data.

I would like to thank my colleagues, Dr. Yeo Jin Kim, Dr. Seung Gu Park, Dr. Changjae Kim, Yeonkyung Kim, Yeonghee Kang, Yeongju Woo, Kyungeun Hong, Jungae Shim, Nayeong Kim, Dr. Hui-Su Kim, Dr. Youngjune Bhak, Yeonsu Jeon, Jasmin Junseo Lee, Yeonsong Choi, Dr. HyoJung Ryu, Chanhan Yoon, Juyeon Park, SangRyoul Han, Suji Hong, Seon Ju Kim, Seobyoon Shin, Hyogyeong Yu, Sukyeon Kim, Asta Blazyte, Jae-Pil Choi, Dr. Byoung-Chul Kim, Yeshin Park, Yumi Kim, Subin Choi, Ji-Hye Ahn, Suan Cho, Dr. Dan Bolser, and those whom I cannot list up, for their support, encouragement, and scientific collaboration. I was lucky to have and work with such great colleagues in my life.

I especially thank my friends Dohyeon Kim, Donghyun Kim, Jae Min Lee, Jongchan Yoon, Kangseok Lee, and Soyoung Park for their encouragement. It has always been happy and lucky to have them during my Ph.D. course.

I thank Wikipedia and Google for helping me to search for various information through the internet.

Finally, I would like to thank my parents and brother for their love and full supports.

This work was supported by the U-K BRAND Research Fund (1.190007.01) of UNIST; Research Project Funded by Ulsan City Research Fund (1.190033.01) of UNIST; Research Project Funded by Ulsan City Research Fund (1.200047.01) of UNIST; Research Project Funded by Ulsan City Research Fund (2.180016.01) of UNIST. This work was also supported by the Technology Innovation Program

(20003641, Development and Dissemination on National Standard Reference Data) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea). This work was also supported by internal funding of Clinomics Inc. Ulsan university hospital biobank provided DNA sample and clinical data from 696 participants of Korea1K (60SA2016001-002, 60SA2016001-003, 60SA2016001-005, 60SA2017002-001, 60SA2017002-004). The Korea Institute of Science and Technology Information (KISTI) provided us the Korea Research Environment Open NETwork (KREONET).

Appendix

Korea1K

SCIENCE ADVANCES | RESEARCH ARTICLE

HUMAN GENETICS

Korean Genome Project: 1094 Korean personal genomes with clinical information

Sungwon Jeon[1,2]*, Youngjune Bhak[1,2,3]*, Yeonsong Choi[1,2]*, Yeonsu Jeon[1,2], Seunghoon Kim[1,2], Jaeyoung Jang[1], Jinho Jang[1,2], Asta Blazyte[1], Changjae Kim[1,3], Yeonkyung Kim[1], Jungae Shim[1], Nayeong Kim[1], Yeo Jin Kim[1], Seung Gu Park[1], Jungeun Kim[4], Yun Sung Cho[3], Yeshin Park[3], Hak-Min Kim[1,2,3], Byoung-Chul Kim[3], Neung-Hwa Park[5,6], Eun-Seok Shin[7], Byung Chul Kim[3], Dan Bolser[3], Andrea Manica[8], Jeremy S. Edwards[9], George Church[10†], Semin Lee[1,2†], Jong Bhak[1,2,3,4†]

We present the initial phase of the Korean Genome Project (Korea1K), including 1094 whole genomes (sequenced at an average depth of 31×), along with data of 79 quantitative clinical traits. We identified 39 million single-nucleotide variants and indels of which half were singleton or doubleton and detected Korean-specific patterns based on several types of genomic variations. A genome-wide association study illustrated the power of whole-genome sequences for analyzing clinical traits, identifying nine more significant candidate alleles than previously reported from the same linkage disequilibrium blocks. Also, Korea1K, as a reference, showed better imputation accuracy for Koreans than the 1KGP panel. As proof of utility, germline variants in cancer samples could be filtered out more effectively when the Korea1K variome was used as a panel of normals compared to non-Korean variome sets. Overall, this study shows that Korea1K can be a useful genotypic and phenotypic resource for clinical and ethnogenetic studies.

INTRODUCTION

The Korean population [estimated census population size close to 85 million (M)] has been thought to be highly homogeneous with few large-scale admixture events in the past (*1*–*4*). However, little formal scrutiny has been given to these claims. Several Korean whole genomes and exomes (*5*, *6*) have been reported since the first Korean genome data (SJK) were published in 2008 (*7*), including the first Korean reference genome sequence (KOREF_S) (*8*) and 40 unrelated individuals (KOREF_C) that formed the basis of KoVariome, the Korean genomic variation database (*9*). Before the current study, at least 100 whole genomes of Korean individuals were available worldwide (*5*, *10*). However, although a global whole-genome project (the multiethnicity 1000 genome project) that aims to characterize global human genetic diversity contains over 2500 genomes, including Chinese and Japanese, it does not include Korean samples yet (*11*).

There has also been an effort to generate ethnicity-specific reference genome sequences, and several human variomes have been generated to expand the coverage of human genome diversity, including the UK10K (*12*), the Genome of the Netherlands (GoNL) project (*13*), and the pan-African genome (*14*). In 2015, the consequences of strong founder effects were demonstrated in the Icelandic population by sequencing 2636 genomes (*15*). In the Danish population study, 150 trios were used to de novo assemble a reference genome, and they provide detailed data on structural variations and many complex genomic regions, including the major histocompatibility complex and major regions of the Y chromosome (*16*). In East Asia, the 1KJPN project yielded data on 1070 Japanese genomes (*17*), and another recent dataset identified selection signatures in the Japanese population from 2234 Japanese whole-genome data (*18*). In contrast, the original KoVariome database contained only 50 Korean whole-genome sequences without clinical information at the time of publication (*9*), although its sample size has subsequently increased to >100 genomes. Despite these large genome sequencing projects in numerous populations, little biochemical and clinical data and limited information regarding genotype-phenotype association for the participants have been collected to characterize the population's health and disease states.

Here, we introduce a dataset comprising 1094 Korean whole genomes of which 1007 genomes were newly generated in combination with systematically acquired clinical and biochemical measurement information from the blood and urine of the participants. This Korea1K set represents the first-phase release of the Korean Genome Project (KGP). KGP is a joint project by the Personal Genome Project at Harvard Medical School, the National Center for Standard Reference Data of Korea, Clinomics Inc., and the Korean Genomics Center of Ulsan National Institute of Science and Technology (UNIST). These genomes have been sequenced to a high sequencing depth (~31× on average) using Illumina HiSeq X10, and we used these data to characterize single-nucleotide variants (SNVs), indels, copy number variations (CNVs), transposable element (TE) insertion, and human leukocyte antigen (HLA) type in the Korean population and contrast the Korean data with similar data from other populations. The majority of the genomic data (984 samples)

[1]Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea. [2]Department of Biomedical Engineering, School of Life Sciences, UNIST, Ulsan 44919, Republic of Korea. [3]Clinomics Inc., Ulsan 44919, Republic of Korea. [4]Personal Genomics Institute (PGI), Genome Research Foundation (GRF), Osong 28160, Republic of Korea. [5]Department of Internal Medicine, University of Ulsan College of Medicine, Ulsan University Hospital, Ulsan 44033, Republic of Korea. [6]Biomedical Research Center, University of Ulsan College of Medicine, Ulsan University Hospital, Ulsan 44033, Republic of Korea. [7]Division of Cardiology, Department of Internal Medicine, Ulsan Medical Center, Ulsan 44686, Republic of Korea. [8]Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK. [9]Department of Chemistry and Chemical Biology, University of New Mexico and University of New Mexico Comprehensive Cancer Center, Albuquerque, NM 87106, USA. [10]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: gchurch@genetics.med.harvard.edu (G.C.); seminlee@unist.ac.kr (S.L.); jongbhak@genomics.org (J.B.)

The origin of Korean

GBE

The Origin and Composition of Korean Ethnicity Analyzed by Ancient and Present-Day Genome Sequences

Jungeun Kim[1,†], Sungwon Jeon[2,3,†], Jae-Pil Choi[1], Asta Blazyte[2], Yeonsu Jeon[2,3], Jong-Il Kim[4], Jun Ohashi[5], Katsushi Tokunaga[6], Sumio Sugano[7], Suthat Fucharoen[8], Fahd Al-Mulla[9], and Jong Bhak[1,2,3,10,*]

[1]Personal Genomics Institute (PGI), Genome Research Foundation, Osong, Republic of Korea

[2]Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea

[3]Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea

[4]Department of Archaeology and Art History, Seoul National University, Republic of Korea

[5]Department of Biological Sciences, Graduate School of Medicine, The University of Tokyo, Japan

[6]Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Japan

[7]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Japan

[8]Thalassemia Research Center, Institute of Molecular Biosciences, Mahidol University, Nakorn Pathom, Thailand

[9]Center of Genomic Medicine, Kuwait University, Kuwait

[10]Clinomics Inc, Ulsan, Republic of Korea

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: jongbhak@genomics.org.

Accepted: March 23, 2020

Data deposition: This project has been deposited at GenBank under the accession provided in supplementary table S1, Supplementary Material online.

Abstract

Koreans are thought to be an ethnic group of admixed northern and southern subgroups. However, the exact genetic origins of these two remain unclear. In addition, the past admixture is presumed to have taken place on the Korean peninsula, but there is no genomic scale analysis exploring the origin, composition, admixture, or the past migration of Koreans. Here, 88 Korean genomes compared with 91 other present-day populations showed two major genetic components of East Siberia and Southeast Asia. Additional paleogenomic analysis with 115 ancient genomes from Pleistocene hunter-gatherers to Iron Age farmers showed a gradual admixture of Tianyuan (40 ka) and Devil's gate (8 ka) ancestries throughout East Asia and East Siberia up until the Neolithic era. Afterward, the current genetic foundation of Koreans may have been established through a rapid admixture with ancient Southern Chinese populations associated with Iron Age Cambodians. We speculate that this admixing trend initially occurred mostly outside the Korean peninsula followed by continuous spread and localization in Korea, corresponding to the general admixture trend of East Asia. Over 70% of extant Korean genetic diversity is explained to be derived from such a recent population expansion and admixture from the South.

Key words: Korean origin, Korean migration, population study, paleogenomics, variome, KoVariome.

Introduction

The 1000 Genome Project (1KGP) showed that East Asians displayed a common genetic bottleneck with non-African humans around the last glacial maximum (1000 Genomes Project Consortium et al. 2015). However, the 1KGP project includes only five EA populations failing to fully represent EA genome structures. In 2009, the HUGO Pan-Asian Consortium (PASNP) confirmed a general concordance between linguistic and genetic affiliations (HUGO Pan-Asian SNP Consortium et al. 2009). Most

Neolithic East Asian ancient genomes

SCIENCE ADVANCES | RESEARCH ARTICLE

EVOLUTIONARY GENETICS

Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago

Veronika Siska,[1]* Eppie Ruth Jones,[1,2] Sungwon Jeon,[3] Youngjune Bhak,[3] Hak-Min Kim,[3] Yun Sung Cho,[3] Hyunho Kim,[4] Kyusang Lee,[5] Elizaveta Veselovskaya,[6] Tatiana Balueva,[6] Marcos Gallego-Llorente,[1] Michael Hofreiter,[7] Daniel G. Bradley,[2] Anders Eriksson,[1] Ron Pinhasi,[8]*† Jong Bhak,[3,4]*†‡ Andrea Manica[1]*†

Ancient genomes have revolutionized our understanding of Holocene prehistory and, particularly, the Neolithic transition in western Eurasia. In contrast, East Asia has so far received little attention, despite representing a core region at which the Neolithic transition took place independently ~3 millennia after its onset in the Near East. We report genome-wide data from two hunter-gatherers from Devil's Gate, an early Neolithic cave site (dated to ~7.7 thousand years ago) located in East Asia, on the border between Russia and Korea. Both of these individuals are genetically most similar to geographically close modern populations from the Amur Basin, all speaking Tungusic languages, and, in particular, to the Ulchi. The similarity to nearby modern populations and the low levels of additional genetic material in the Ulchi imply a high level of genetic continuity in this region during the Holocene, a pattern that markedly contrasts with that reported for Europe.

INTRODUCTION

Ancient genomes from western Asia have revealed a degree of genetic continuity between preagricultural hunter-gatherers and early farmers 12 to 8 thousand years ago (ka) (*1*, *2*). In contrast, studies on southeast and central Europe indicate a major population replacement of Mesolithic hunter-gatherers by Neolithic farmers of a Near Eastern origin during the period 8.5 to 7 ka. This is then followed by a progressive "resurgence" of local hunter-gatherer lineages in some regions during the Middle/Late Neolithic and Eneolithic periods and a major contribution from the Asian Steppe later, ~5.5 ka, coinciding with the advent of the Bronze Age (*3*–*5*). Compared to western Eurasia, for which hundreds of partial ancient genomes have already been sequenced, East Asia has been largely neglected by ancient DNA studies to date, with the exception of the Siberian Arctic belt, which has received attention in the context of the colonization of the Americas (*6*, *7*). However, East Asia represents an extremely interesting region as the shift to reliance on agriculture appears to have taken a different course from that in western Eurasia. In the latter region, pottery, farming, and animal husbandry were closely associated. In contrast, Early Neolithic societies in the Russian Far East, Japan, and Korea started to manufacture and use pottery and basketry 10.5 to 15 ka, but domesticated crops and livestock arrived several millennia later (*8*, *9*). Because of the current lack of ancient genomes from East Asia, we do not know the extent to which this gradual Neolithic transition, which happened independently from the one taking place in western Eurasia, reflected actual migrations, as found in Europe, or the cultural diffusion associated with population continuity.

[1]Department of Zoology, University of Cambridge, Downing Street, Cambridge CB23EJ, U.K. [2]Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland. [3]The Genomics Institute, Ulsan National Institute of Science and Technology, Ulsan 44919, Republic of Korea. [4]Geromics, Ulsan 44919, Republic of Korea. [5]Clinomics Inc., Ulsan 4919, Republic of Korea. [6]Institute of Ethnology and Anthropology, Russian Academy of Sciences, Moscow, Russia. [7]Institute for Biochemistry and Biology, Faculty for Mathematics and Natural Sciences, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam-Golm, Germany. [8]School of Archaeology and Earth Institute, University College Dublin, Dublin, Ireland.
*Corresponding author. Email: vs389@cam.ac.uk (V.S.); ron.pinhasi@ucd.ie (R.P.); jongbhak@genomics.org (J.B.); am315@cam.ac.uk (A.M.)
†These authors contributed equally to this work.
‡Adjunct professor at Seoul National University, Seoul, Republic of Korea.

RESULTS

Samples, sequencing, and authenticity

To fill this gap in our knowledge about the Neolithic in East Asia, we sequenced to low coverage the genomes of five early Neolithic burials (DevilsGate1, 0.059-fold coverage; DevilsGate2, 0.023-fold coverage; and DevilsGate3, DevilsGate4, and DevilsGate5, <0.001-fold coverage) from a single occupational phase at Devil's Gate (Chertovy Vorota) Cave in the Primorye Region, Russian Far East, close to the border with China and North Korea (see the Supplementary Materials). This site dates back to 9.4 to 7.2 ka, with the human remains dating to ~7.7 ka, and it includes some of the world's earliest evidence of ancient textiles (*10*). The people inhabiting Devil's Gate were hunter-fisher-gatherers with no evidence of farming; the fibers of wild plants were the main raw material for textile production (*10*). We focus our analysis on the two samples with the highest sequencing coverage, DevilsGate1 and DevilsGate2, both of which were female. The mitochondrial genome of the individual with higher coverage (DevilsGate1) could be assigned to haplogroup D4; this haplogroup is found in present-day populations in East Asia (*11*) and has also been found in Jomon skeletons in northern Japan (*2*). For the other individual (DevilsGate2), only membership to the M branch (to which D4 belongs) could be established. Contamination, estimated from the number of discordant calls in the mitochondrial DNA (mtDNA) sequence, was low {0.87% [95% confidence interval (CI), 0.28 to 2.37%] and 0.59% (95% CI, 0.03 to 3.753%)} on nonconsensus bases at haplogroup-defining positions for DevilsGate1 and DevilsGate2, respectively. Using schmutzi (*12*) on the higher-coverage genome, DevilsGate1 also gives low contamination levels [1% (95% CI, 0 to 2%); see the Supplementary Materials]. As a further check against the possible confounding effect of contamination, we made sure that our most important analyses [outgroup f_3 scores and principal components analysis (PCA)] were qualitatively replicated using only reads showing evidence of postmortem damage (PMD score of at least 3) (*13*), although these latter results had a high level of noise due to the low coverage (0.005*X* for DevilsGate1 and 0.001*X* for DevilsGate2).

Welfare genome project

frontiers in Genetics

PERSPECTIVE
published: 09 February 2021
doi: 10.3389/fgene.2021.633731

Welfare Genome Project: A Participatory Korean Personal Genome Project With Free Health Check-Up and Genetic Report Followed by Counseling

Yeonsu Jeon[1,2†], Sungwon Jeon[1,2†], Asta Blazyte[1,2†], Yeo Jin Kim[3], Jasmin Junseo Lee[1,4], Youngjune Bhak[1,2], Yun Sung Cho[3], Yeshin Park[3,5], Eui-Kyu Noh[6], Andrea Manica[7], Jeremy S. Edwards[8], Dan Bolser[9], Sukyeon Kim[1], Yuji Lee[1], Changhan Yoon[1,2], Semin Lee[1,2], Byung Chul Kim[3], Neung Hwa Park[10]* and Jong Bhak[1,2,3,11]*

[1] Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea, [2] Department of Biomedical Engineering, College of Information-Bio Convergence Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea, [3] Clinomics Inc., Ulsan, South Korea, [4] Human Biology Program, Faculty of Arts and Sciences, University of Toronto, Toronto, ON, Canada, [5] Department of Medical Sciences, Graduate School of Ajou University School, Suwon, South Korea, [6] Department of Hematology and Oncology, Ulsan University Hospital, University of Ulsan College of Medicine, Ulsan, South Korea, [7] Department of Zoology, University of Cambridge, Cambridge, United Kingdom, [8] Department of Chemistry and Chemical Biology, University of New Mexico Comprehensive Cancer Center, University of New Mexico, Albuquerque, NM, United States, [9] Geromics Ltd., Cambridge, United Kingdom, [10] Department of Internal Medicine, Ulsan University Hospital, University of Ulsan College of Medicine, Ulsan, South Korea, [11] Personal Genomics Institute (PGI), Genome Research Foundation (GRF), Osong, South Korea

The Welfare Genome Project (WGP) provided 1,000 healthy Korean volunteers with detailed genetic and health reports to test the social perception of integrating personal genetic and healthcare data at a large-scale. WGP was launched in 2016 in the Ulsan Metropolitan City as the first large-scale genome project with public participation in Korea. The project produced a set of genetic materials, genotype information, clinical data, and lifestyle survey answers from participants aged 20–96. As compensation, the participants received a free general health check-up on 110 clinical traits, accompanied by a genetic report of their genotypes followed by genetic counseling. In a follow-up survey, 91.0% of the participants indicated that their genetic reports motivated them to improve their health. Overall, WGP expanded not only the general awareness of genomics, DNA sequencing technologies, bioinformatics, and bioethics regulations among all the parties involved, but also the general public's understanding of how genome projects can indirectly benefit their health and lifestyle management. WGP established a data construction framework for not only scientific research but also the welfare of participants. In the future, the WGP framework can help lay the groundwork for a new personalized healthcare system that is seamlessly integrated with existing public medical infrastructure.

Keywords: genomics, personal genome project, Korean genome project, population study, integrated healthcare, genetic report

OPEN ACCESS

Edited by:
Go Yoshizawa, Kwansei Gakuin University, Japan

Reviewed by:
Juliana F. Mazzeu, University of Brasília, Brazil
Giovanna Elisa Calabrò, Catholic University of the Sacred Heart, Italy

***Correspondence:**
Neung Hwa Park
nhparkmd@gmail.com
Jong Bhak
jongbhak@genomics.org

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Genetics

Received: *27 November 2020*
Accepted: *20 January 2021*
Published: *09 February 2021*

Citation:
Jeon Y, Jeon S, Blazyte A, Kim YJ, Lee JJ, Bhak Y, Cho YS, Park Y, Noh EK, Manica A, Edwards JS, Bolser D, Kim S, Lee Y, Yoon C, Lee S, Kim BC, Park NH and Bhak J (2021) Welfare Genome Project: A Participatory Korean Personal Genome Project With Free Health Check-Up and Genetic Report Followed by Counseling. Front. Genet. 12:633731. doi: 10.3389/fgene.2021.633731

Korean Reference genome assembly with PromethION, PacBio, and Hi-C

OXFORD (GIGA)n SCIENCE

GigaScience, 8, 2019, 1–5

doi: 10.1093/gigascience/giz125
Data Note

DATA NOTE

Chromosome-scale assembly comparison of the Korean Reference Genome KOREF from PromethION and PacBio with Hi-C mapping information

Hui-Su Kim[1], Sungwon Jeon[1,2], Changjae Kim[1], Yeon Kyung Kim[1], Yun Sung Cho[3], Jungeun Kim[4], Asta Blazyte[1], Andrea Manica[5], Semin Lee[1,2,*] and Jong Bhak[1,2,3,4,*]

[1]KOGIC, Ulsan National Institute of Science and Technology (UNIST), UNIST-gil 50, Eonyang-eup, Ulju-gun, Ulsan 44919, Republic of Korea; [2]Department of Biomedical Engineering, School of Life Sciences, UNIST-gil 50, Eonyang-eup, Ulju-gun, UNIST, Ulsan 44919, Republic of Korea; [3]Clinomics Inc., UNIST-gil 50, Eonyang-eup, Ulju-gun, Ulsan 44919, Republic of Korea; [4]Personal Genomics Institute, Genome Research Foundation, Osong saengmyong1ro, Cheongju 28160, Republic of Korea and [5]Department of Zoology, Cambridge University, Downing street, Cambridge CB2 3EJ, UK

*Correspondence address. Semin Lee, #110-302, Ulsan National Institute of Science and Technology, UNIST-gil 50, Eonyang-eup, Ulju-gun, Ulsan 44919, Republic of Korea. Tel: +82 52-217-2663; E-mail: seminlee@gmail.com http://orcid.org/0000-0002-9015-6046; Jong Bhak, #110-303, Ulsan National Institute of Science and Technology, UNIST-gil 50, Eonyang-eup, Ulju-gun, Ulsan 44919, Republic of Korea. Tel: +82 10-4644-6754; E-mail: jongbhak@genomics.org http://orcid.org/0000-0002-4228-1299

Abstract

Background: Long DNA reads produced by single-molecule and pore-based sequencers are more suitable for assembly and structural variation discovery than short-read DNA fragments. For *de novo* assembly, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are the favorite options. However, PacBio's SMRT sequencing is expensive for a full human genome assembly and costs more than $40,000 US for 30× coverage as of 2019. ONT PromethION sequencing, on the other hand, is 1/12 the price of PacBio for the same coverage. This study aimed to compare the cost-effectiveness of ONT PromethION and PacBio's SMRT sequencing in relation to the quality. **Findings:** We performed whole-genome *de novo* assemblies and comparison to construct an improved version of KOREF, the Korean reference genome, using sequencing data produced by PromethION and PacBio. With PromethION, an assembly using sequenced reads with 64× coverage (193 Gb, 3 flowcell sequencing) resulted in 3,725 contigs with N50s of 16.7 Mb and a total genome length of 2.8 Gb. It was comparable to a KOREF assembly constructed using PacBio at 62× coverage (188 Gb, 2,695 contigs, and N50s of 17.9 Mb). When we applied Hi-C–derived long-range mapping data, an even higher quality assembly for the 64× coverage was achieved, resulting in 3,179 scaffolds with an N50 of 56.4 Mb. **Conclusion:** The pore-based PromethION approach provided a high-quality chromosome-scale human genome assembly at a low cost with long maximum contig and scaffold lengths and was more cost-effective than PacBio at comparable quality measurements.

Keywords: Korean reference genome; KOREF; PromethION; Hi-C; nanopore sequencing; single-molecule sequencing

Received: 27 June 2019; **Revised:** 2 September 2019; **Accepted:** 28 September 2019

Korean early-onset acute myocardial infarction genome analysis

PLOS ONE

RESEARCH ARTICLE

Polygenic risk score validation using Korean genomes of 265 early-onset acute myocardial infarction patients and 636 healthy controls

Youngjune Bhak[1,2☯], Yeonsu Jeon[1,2☯], Sungwon Jeon[1,2☯], Changhan Yoon[1,2], Min Kim[1,2], Asta Blazyte[1,2], Yeonkyung Kim[1], Younghui Kang[1], Changjae Kim[3], Sang Yeub Lee[4], Jang-Whan Bae[4], Weon Kim[5], Yeo Jin Kim[1], Jungae Shim[1], Nayeong Kim[1], Sung Chun[6,7], Byoung-Chul Kim[3], Byung Chul Kim[3], Semin Lee[1,2], Jong Bhak[1,2,3,8]*, Eun-Seok Shin[8,9]*

1 Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea, **2** Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea, **3** Clinomics Inc, Ulsan, Republic of Korea, **4** Division of Cardiology, Department of Internal Medicine, Chungbuk National University, College of Medicine, Cheongju, Republic of Korea, **5** Division of Cardiology, Department of Internal Medicine, Kyung Hee University Hospital, Seoul, Republic of Korea, **6** Division of Pulmonary Medicine, Boston Children's Hospital, Boston, Massachusetts, United States of America, **7** Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, United States of America, **8** Personal Genomics Institute, Genome Research Foundation, Ulsan, Republic of Korea, **9** Division of Cardiology, Department of Internal Medicine, Ulsan Medical Center, Ulsan, Republic of Korea

☯ These authors contributed equally to this work.
* jongbhak@genomics.org (JB); sesim1989@gmail.com (ESS)

OPEN ACCESS

Citation: Bhak Y, Jeon Y, Jeon S, Yoon C, Kim M, Blazyte A, et al. (2021) Polygenic risk score validation using Korean genomes of 265 early-onset acute myocardial infarction patients and 636 healthy controls. PLoS ONE 16(2): e0246538. https://doi.org/10.1371/journal.pone.0246538

Editor: Yiqiang Zhan, German Centre for Neurodegenerative Diseases Site Munich: Deutsches Zentrum fur Neurodegenerative Erkrankungen Standort Munchen, GERMANY

Received: September 1, 2020

Accepted: January 21, 2021

Published: February 4, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0246538

Data Availability Statement: The data presented is under legal restriction since data contain potentially

Abstract

Background

The polygenic risk score (PRS) developed for coronary artery disease (CAD) is known to be effective for classifying patients with CAD and predicting subsequent events. However, the PRS was developed mainly based on the analysis of Caucasian genomes and has not been validated for East Asians. We aimed to evaluate the PRS in the genomes of Korean early-onset AMI patients (n = 265, age ≤50 years) following PCI and controls (n = 636) to examine whether the PRS improves risk prediction beyond conventional risk factors.

Results

The odds ratio of the PRS was 1.83 (95% confidence interval [CI]: 1.69–1.99) for early-onset AMI patients compared with the controls. For the classification of patients, the area under the curve (AUC) for the combined model with the six conventional risk factors (diabetes mellitus, family history of CAD, hypertension, body mass index, hypercholesterolemia, and current smoking) and PRS was 0.92 (95% CI: 0.90–0.94) while that for the six conventional risk factors was 0.91 (95% CI: 0.85–0.93). Although the AUC for PRS alone was 0.65 (95% CI: 0.61–0.69), adding the PRS to the six conventional risk factors significantly improved the accuracy of the prediction model ($P = 0.015$). Patients with the upper 50% of PRS showed a higher frequency of repeat revascularization (hazard ratio = 2.19, 95% CI: 1.47–3.26) than the others.

Egyptian genomes

Gene 668 (2018) 129–134

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Whole genome sequencing and bioinformatics analysis of two Egyptian genomes

Mahmoud ElHefnawi[a,*,1], Sungwon Jeon[b,c,1], Youngjune Bhak[b,c], Asmaa ElFiky[a,d], Ahmed Horaiz[a], JeHoon Jun[e,f], Hyunho Kim[f], Jong Bhak[b,c,e,f,**]

[a] Biomedical Informatics and Chemo-Informatics Group, Centre of Excellence for Advanced Sciences (CEAS), and Informatics and Systems Department, National Research Centre, Cairo 12622, Egypt
[b] Korean Genomics Industrialization and Commercialization Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea
[c] Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea.
[d] Environmental and Occupational Medicine Department, National Research Centre, Cairo 12622, Egypt
[e] Personal Genomics Institute, Genome Research Foundation, Cheongju 28160, Republic of Korea
[f] Geromics, Ulsan 44919, Republic of Korea.

ARTICLE INFO

Keywords:
Whole-genome sequencing
Egyptian
Variants
Human migration
Bioinformatics

ABSTRACT

We report two Egyptian male genomes (EGP1 and EGP2) sequenced at ~ 30× sequencing depths. EGP1 had 4.7 million variants, where 198,877 were novel variants while EGP2 had 209,109 novel variants out of 4.8 million variants. The mitochondrial haplogroup of the two individuals were identified to be H7b1 and L2a1c, respectively. We also identified the Y haplogroup of EGP1 (R1b) and EGP2 (J1a2a1a2 > P58 > FGC11). EGP1 had a mutation in the NADH gene of the mitochondrial genome ND4 (m.11778 G > A) that causes Leber's hereditary optic neuropathy. Some SNPs shared by the two genomes were associated with an increased level of cholesterol and triglycerides, probably related with Egyptians obesity. Comparison of these genomes with African and Western-Asian genomes can provide insights on Egyptian ancestry and genetic history. This resource can be used to further understand genomic diversity and functional classification of variants as well as human migration and evolution across Africa and Western-Asia.

1. Introduction

A human genome holds an extensive amount of data on human evolution, diversity, health, physiology, and medicine (Lander et al., 2001). Whole genome sequencing (WGS) data can be used for the deepest possible genetic analyses for various purposes such as common and rare disorder association studies. Genomes and their diverse variation information can also be used effectively for estimating risk factors of common diseases (Bick & Dimmock, 2011; Thompson et al., 2012). Currently, massively-parallel next-generation sequencing (NGS) methods are the most widely used method for analyzing the whole human genomes. Programs to map short reads of a genome and to call the subsequent variations are being rapidly improved and upgraded (Lupski et al., 2010). In addition, the cost of analyzing a genome has become very low and WGS is becoming more common in detecting uncommon, disease-causing variants by scrutinizing affected people's genomes (Lupski et al., 2010; Sobreira et al., 2010; Roach et al., 2010). For example, it can be useful for screening women who have BRCA1 and BRCA2 genes mutations to assess the risk of breast and ovarian cancers (Campeau et al., 2008).

The Egyptian population is diverse due to its position between Africa and Asia. It has two long banks along the Nile River, which is the longest African River, and has hosted various populations throughout history. Ancient Egyptian traditions, such as mummification, play an important role in preserving genomes and subsequent analysis of DNA variants (Paabo, 1985). Egyptian DNA have been studied for a long time

Abbreviations: EGP, Egyptian person; HQ, High quality; LD, Linkage disequilibrium; LHON, Leber's hereditary optic neuropathy; mtDNA, Mitochondrial DNA; NGS, Next generation sequencing; rCRS, revised Cambridge reference sequence; SNP, Single nucleotide polymorphism; WGS, Whole genome sequencing; Y-STR, short tandem repeat (STR) on the Y-chromosome

* Corresponding author.
** Correspondence to: J. Bhak, Korean Genomics Industrialization and Commercialization Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea.
E-mail addresses: mahef@aucegypt.edu (M. ElHefnawi), jongbhak@genomics.org (J. Bhak).
[1] Equal contributors.

https://doi.org/10.1016/j.gene.2018.05.048
Received 28 March 2018; Accepted 13 May 2018
Available online 25 May 2018
0378-1119/

Nine Egyptian genomes

Gene 769 (2021) 145237

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Complete genome sequence and bioinformatics analysis of nine Egyptian females with clinical information from different geographic regions in Egypt

Mahmoud ElHefnawi [a,b,c,1,*], Elsayed Hegazy [a,1], Asmaa Elfiky [d,1], Yeonsu Jeon [e,f], Sungwon Jeon [e,f], Jong Bhak [e,f,g], Fateheya Mohamed Metwally [d], Sumio [illegible] [h], Terumi Horiuchi [i], Abe Kazumi [h], Asta Blazyte [e,f,1]

[a] School of Information Technology and Computer Science, Nile University, Giza 12588, Egypt
[b] Informatics & Systems Department, the National Research Centre, Cairo, Egypt
[c] Biomedical Informatics and Chemoinformatics Group, Center of Excellence for Medical Research, National Research Centre, Cairo, Egypt
[d] Environmental and Occupational Medicine Department, Environmental Research Division, National Research Centre, Cairo, Egypt
[e] Korean Genomics Center (KOGIC), UNIST, Republic of Korea
[f] Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea
[g] Personal Genomics Institute, Genome Research Foundation, Osong, Republic of Korea
[h] The Institute of Medical Science, University of Tokyo, Japan
[i] Graduate School of Frontier Sciences, University of Tokyo, Chiba, Japan

Go to page 1

ARTICLE INFO

Original content: https://www.ncbi.nlm.nih.gov/sra/?term=SRP136979

Keywords:
Personal genome
Variant analysis

ABSTRACT

Egyptians are at a crossroad between Africa and Eurasia, providing useful genomic resources for analyzing both genetic and environmental factors for future personalized medicine. Two personal Egyptian whole genomes have been published previously by us and here nine female whole genome sequences with clinical information have been added to expand the genomic resource of Egyptian personal genomes. Here we report the analysis of whole genomes of nine Egyptian females from different regions using Illumina short-read sequencers. At 30x sequencing coverage, we identified 12 SNPs that were shared in most of the subjects associated with obesity which are concordant with their clinical diagnosis. Also, we found mtDNA mutation A4282G is common in all the samples and this is associated with chronic progressive external ophthalmoplegia (CPEO). Haplogroup and Admixture analyses revealed that most Egyptian samples are close to the other north Mediterranean, Middle Eastern, and European, respectively, possibly reflecting the into-Africa influx of human migration. In conclusion, we present whole-genome sequences of nine Egyptian females with personal clinical information that cover the diverse regions of Egypt. Although limited in sample size, the whole genomes data provides possible geno-phenotype candidate markers that are relevant to the region's diseases.

1. Introduction

Next-generation sequencing (NGS) technology is a powerful approach enabling an efficient study of a large scale genetic variants (Bentley et al., 2008; Wheeler et al., 2008; Metzker, 2010; Azim et al., 2013). Whole genome sequencing (WGS) by NGS can cover intronic areas that may contain rare and common deleterious mutations, that cannot be captured by whole exome sequencing (WES) that usually facilitates a deeper coverage of coding regions important for protein function analyses (Bamshad et al., 2011). One of the emerging applications of NGS is investigating complex diseases such as obesity along with its manifestations: insulin resistance, impaired glucose tolerance, and dyslipidemia. Even though Genome-Wide Association Studies (GWAS) may rely on genotyping data like in Frayling et al. (2007) and

Abbreviations: WGS, Whole-genome sequencing; CPEO, chronic progressive external ophthalmoplegia; NGS, next-generation sequencing; GWAS, Genome-Wide Association Studies; EGY, Egyptian; VEST, Variant Effect Scoring Tool; rCRS, Revised Cambridge Reference Sequence.

* Corresponding author at: Informatics & Systems Department, the National Research Centre, Cairo, Egypt.
E-mail address: mahef@aucegypt.edu (M. ElHefnawi).

[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.gene.2020.145237
Received 25 November 2019; Received in revised form 3 August 2020; Accepted 11 October 2020
Available online 27 October 2020

Indian genome

Genomics 111 (2019) 196–204

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Sequencing and analysis of the whole genome of Indian Gujarati male

Suhani Almal[b], Sungwon Jeon[c], Milee Agarwal[b], Sweta Patel[b], Shivangi Patel[b], Youngjune Bhak[c], JeHoon Jun[d], Jong Bhak[c,d,e], Harish Padh[a,*]

[a] Sardar Patel University, Vallabh Vidyanagar, Gujarat, India
[b] B. V. Patel Pharmaceutical Education and Research Development (PERD) Centre, Thaltej, Ahmedabad, Gujarat, India
[c] The Genomics Institute, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea
[d] Personal Genomics Institute, Genome Research Foundation, Cheongju 28160, Republic of Korea
[e] Geromics, Ulsan 44919, Republic of Korea

ARTICLE INFO

Keywords:
Whole-genome sequencing
Gujarati Indian male
Variant analysis
Health profile
Human migration

ABSTRACT

The article presents the analysis of whole genome sequence of a Gujarati Indian individual (IHGP01) that was sequenced at 23.05× coverage with a total of 74.93 Gb of sequence data generated using Illumina HiSeq 2000 platform. Variant analysis revealed over 3.9 million single nucleotide variants (SNVs) and about 393,000 small insertions and deletions (InDels) including novel variants. The known variants were analyzed for their health and disease relevance and pharmacogenomic profile. Mitochondrial and Y-chromosome haplogroup analysis clearly indicated arrival on the continent not more than 20,000–25,000 years ago, following the route out of Africa to central Europe, then into Asian continent and subsequent migration to West part of the Indian subcontinent. The current research has added 141,000 novel genetic variations to the human DNA database. Functional analysis and validation of these novel variations and revelation of their role in health and disease will add a newer dimension to understand people of this subcontinent.

1. Introduction

Human reference genome originally developed in 2003 is derived from about 200 anonymous individuals from six countries: China, France, Germany, Great Britain, Japan, and the United States [1]. The reference genome provides the framework to which any individual genome can be compared to reveal the individual's genetic make-up. With the technology now being more accessible and affordable and the reference database sufficiently enriched, analyzing individual genome has become feasible to understand one's genetic background for health and associated risk factors. Over a period of time, complete individual genome or selective portions (exome sequencing WES, etc.) have been analyzed. Collectively the reference database has revealed a number of important aspects of the development of the human race. Particularly, Y-chromosome and mitochondrial haplogroups have helped to develop high-resolution human migration map during the past 100,000 years. With individual genomes, we are now able to ascertain the ethnic affiliation of the ancestors and develop a migratory pattern for that particular group and also develop health profile of the individual.

India occupies the central position on migratory routes to South East Asia and with its 1.25 billion people, comprises about one-sixth of the world population. The Indian population is believed to have developed over a period of last 80,000 years through several distinct migratory waves from different directions. The contemporary Indian population is an admixture of such populations, which is believed to have been segregated into various communities only recently over a period of last 1000 years or so [2].

It is to be noted that original Human Reference Genome had no representation from this subcontinent. Although subsequently, 1000 genome project [3,4] had some individuals of Indian origin, overall the current database of human variants is still lacking in proportionate representation from this subcontinent. More variants from Indian subcontinent need to be added to make the reference database as complete representative of Reference Human Database. In 2003, a large-scale study of the genetics of the Indian population was initiated through Indian Genome Variation Consortium Project. In this project,15,000 individuals were screened for over 1000 biomedically important genes [5–7] (database URL; http://www.igvdb.res.in). However, the project scanned for only the reported known variants and did not report the presence of any novel (not reported earlier) variants.

The objectives of the present study are: (1) complete sequencing of an Indian individual and analysis of the genome variants for known risk factors for health and disease; (2) to identify unreported novel variants present in this individual and enrich the human variant database with

* Corresponding author at: Sardar Patel University, Vallabh Vidyanagar-388 120, Gujarat, India.
E-mail address: hpadh@yahoo.com (H. Padh).

https://doi.org/10.1016/j.ygeno.2018.02.003
Received 14 August 2017; Received in revised form 29 January 2018; Accepted 8 February 2018
Available online 10 February 2018

Homotherium latidens ancient genome

CellPress
OPEN ACCESS

Current Biology

Report

Genomic Adaptations and Evolutionary History of the Extinct Scimitar-Toothed Cat, *Homotherium latidens*

Ross Barnett,[1,35] Michael V. Westbury,[1,35,36,*] Marcela Sandoval-Velasco,[1,35] Filipe Garrett Vieira,[1] Sungwon Jeon,[2,3] Grant Zazula,[4] Michael D. Martin,[5] Simon Y.W. Ho,[6] Niklas Mather,[6] Shyam Gopalakrishnan,[1,7] Jazmín Ramos-Madrigal,[1,7] Marc de Manuel,[8] M. Lisandra Zepeda-Mendoza,[1,9] Agostinho Antunes,[10,11] Aldo Carmona Baez,[1] Binia De Cahsan,[1] Greger Larson,[12] Stephen J. O'Brien,[13,14] Eduardo Eizirik,[15,32,33]

(Author list continued on next page)

[1]Section for Evolutionary Genomics, The GLOBE Institute, University of Copenhagen, Øster Voldgade 5-7, Copenhagen, Denmark
[2]Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea
[3]Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea
[4]Yukon Palaeontology Program, Department of Tourism and Culture, Government of Yukon, PO Box 2703, Whitehorse, YT Y1A 2C6, Canada
[5]Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology (NTNU), Trondheim NO-7491, Norway
[6]School of Life and Environmental Sciences, University of Sydney, NSW 2006, Australia
[7]Center for Evolutionary Hologenomics, The GLOBE Institute, University of Copenhagen, Øster Farimagsgade 5A, Copenhagen 1352, Denmark
[8]Institute of Evolutionary Biology (UPF-CSIC), PRBB, Dr. Aiguader 88, Barcelona 08003, Spain
[9]School of Medical and Dental Sciences, Institute of Microbiology and Infection, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK
[10]CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos, s/n, Porto 4450-208, Portugal
[11]Department of Biology, Faculty of Sciences, University of Porto, Porto 4169-007, Portugal

(Affiliations continued on next page)

SUMMARY

Homotherium was a genus of large-bodied scimitar-toothed cats, morphologically distinct from any extant felid species, that went extinct at the end of the Pleistocene [1–4]. They possessed large, saber-form serrated canine teeth, powerful forelimbs, a sloping back, and an enlarged optic bulb, all of which were key characteristics for predation on Pleistocene megafauna [5]. Previous mitochondrial DNA phylogenies suggested that it was a highly divergent sister lineage to all extant cat species [6–8]. However, mitochondrial phylogenies can be misled by hybridization [9], incomplete lineage sorting (ILS), or sex-biased dispersal patterns [10], which might be especially relevant for *Homotherium* since widespread mito-nuclear discrepancies have been uncovered in modern cats [10]. To examine the evolutionary history of *Homotherium*, we generated a ~7x nuclear genome and a ~38x exome from *H. latidens* using shotgun and target-capture sequencing approaches. Phylogenetic analyses reveal *Homotherium* as highly divergent (~22.5 Ma) from living cat species, with no detectable signs of gene flow. Comparative genomic analyses found signatures of positive selection in several genes, including those involved in vision, cognitive function, and energy consumption, putatively consistent with diurnal activity, well-developed social behavior, and cursorial hunting [5]. Finally, we uncover relatively high levels of genetic diversity, suggesting that *Homotherium* may have been more abundant than the limited fossil record suggests [3, 4, 11–14]. Our findings complement and extend previous inferences from both the fossil record and initial molecular studies, enhancing our understanding of the evolution and ecology of this remarkable lineage.

RESULTS AND DISCUSSION

We used a combination of shotgun and whole genome and exome target-capture Illumina sequencing to generate the nuclear genome of a single *Homotherium latidens* individual to a depth of ~7x coverage, and its exome to ~38x coverage. The genome was sequenced from a fossil humerus (specimen YG 439.38), which was determined to be older than the limits of radiocarbon dating (>47.5 kya [UCIAMS-142835]), recovered from Pleistocene permafrost sediments near Dawson City, Yukon Territory, Canada

The greater wax moth genome assembly and multiomics analysis

Cell Reports
Article

OPEN ACCESS
CellPress

The *Galleria mellonella* Hologenome Supports Microbiota-Independent Metabolism of Long-Chain Hydrocarbon Beeswax

Hyun Gi Kong,[1,4] Hyun Ho Kim,[3,4] Joon-hui Chung,[1,4] JeHoon Jun,[3] Soohyun Lee,[1] Hak-Min Kim,[2] Sungwon Jeon,[2] Seung Gu Park,[2] Jong Bhak,[2,3] and Choong-Min Ryu[1,5,*]
[1]Molecular Phytobacteriology Laboratory, Infection Disease Research Center, KRIBB, Daejeon 34141, South Korea
[2]Biomedical Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea
[3]The Clinomics Institute, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea
[4]These authors contributed equally
[5]Lead Contact
*Correspondence: cmryu@kribb.re.kr
https://doi.org/10.1016/j.celrep.2019.02.018

SUMMARY

The greater wax moth, *Galleria mellonella*, degrades wax and plastic molecules. Despite much interest, the genetic basis of these hallmark traits remains poorly understood. Herein, we assembled high-quality genome and transcriptome data from *G. mellonella* to investigate long-chain hydrocarbon wax metabolism strategies. Specific carboxylesterase and lipase and fatty-acid-metabolism-related enzymes in the *G. mellonella* genome are transcriptionally regulated during feeding on beeswax. Strikingly, *G. mellonella* lacking intestinal microbiota successfully decomposes long-chain fatty acids following wax metabolism, although the intestinal microbiome performs a supplementary role in short-chain fatty acid degradation. Notably, final wax derivatives were detected by gas chromatography even in the absence of gut microbiota. Our findings provide insight into wax moth adaptation and may assist in the development of unique wax-degradation strategies with a similar metabolic approach for a plastic molecule polyethylene biodegradation using organisms without intestinal microbiota.

INTRODUCTION

In recent decades, plastics have been routinely released into the environment via sewage treatment plants, waste disposal, and aerial deposition, and global plastic production has expanded tremendously worldwide (Nowack and Bucheli, 2007). Plastic disposal is one of the biggest problems facing the environment, because vast amounts of synthetic plastic remain nondegradable (Nkwachukwu et al., 2013). Plastics are synthetic polymers composed of carbon, hydrogen, oxygen, and chloride that are derived from multiple sources, such as petroleum, coal, and natural gas. The most widely used plastics polymers are polyethylene (PE), polypropylene (PP), PE terephthalate (PET), polystyrene (PS), and polyvinyl chloride (PVC) (Wu et al., 2017). PE, the most common petroleum-based plastic, is widely used in everyday life. However, the high durability and short usage time of PE is resulting in rapid accumulation in the environment, raising international interest (Ammala et al., 2011; Roy et al., 2011; Shah et al., 2008; Zettler et al., 2013).

The potential to decompose plastics in various environments has been studied in order to investigate biological degradation as a solution to accumulating plastics in the environment (Albertsson and Karlsson, 1988; Artham et al., 2009; Jones et al., 1974; Ohtake et al., 1998; Pegram and Andrady, 1989). Biodegradation of PE in the environment occurs mainly through the biological activity of microorganisms after thermal oxidation (Albertsson et al., 1987; Tokiwa et al., 2009). PE is decomposed into low-molecular-weight substances such as alkanes, alkenes, ketones, aldehydes, various alcohols, and fatty acids (Albertsson et al., 1987, 1998; Tokiwa et al., 2009). More than 90 genera of bacteria and fungi have been proposed to possess the ability to break down plastics (Mahdiyah and Mukti, 2013). However, many plastic components are recalcitrant to biodegradation by microorganisms, and the processing capacity is a generally very slow (Singh and Gupta, 2014). Metabolism of long-chain hydrocarbons is the most important step in the biodegradation of PE. This activity has not previously been reported in microorganisms. Interestingly, naturally occurring beeswax is a natural substance consisting of palmitoleate, long-chain aliphatic alcohols, and hydrocarbons. Similarly, PE is composed of a long-chain linear backbone of carbon atoms. The production of long-chain fatty acids and long-chain ethanol from beeswax is the most important process in long-chain hydrocarbon degradation. However, the associated genes and enzymes have not been studied in microorganisms.

Alternatively, the potential to metabolize long-chain hydrocarbons using insects has been studied extensively, because the enzymes and mechanisms mediating the biodegradation of long-chain hydrocarbons in environmental microorganisms remain elusive. However, *Tenebrio molitor* larvae (or mealworms) from a source in Beijing showed PS-degrading capacity, and a gut-PS-degrading *Exiguobacterium* spp. strain YT2 was isolated. The ubiquity of gut-microbiota-dependent PS degradation by mealworms was demonstrated later (Yang et al., 2018a, 2018b). Mealworms can also biodegrade PE (Brandon et al.,

The whale shark genome assembly

The whale shark genome reveals how genomic and physiological properties scale with body size

Jessica A. Weber[a,1], Seung Gu Park[b,c,1], Victor Luria[d,1], Sungwon Jeon[b,c], Hak-Min Kim[b,c], Yeonsu Jeon[b,c], Youngjune Bhak[b,c], Je Hun Jun[e], Sang Wha Kim[f,g], Won Hee Hong[h], Semin Lee[b,c], Yun Sung Cho[e], Amir Karger[i], John W. Cain[j], Andrea Manica[k], Soonok Kim[l], Jae-Hoon Kim[m], Jeremy S. Edwards[n,2,3], Jong Bhak[b,c,e,2,3], and George M. Church[a,2,3]

[a]Department of Genetics, Harvard Medical School, Boston, MA 02115; [b]Korean Genomics Center, Ulsan National Institute of Science and Technology, 44919 Ulsan, Republic of Korea; [c]Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology, 44919 Ulsan, Republic of Korea; [d]Department of Systems Biology, Harvard Medical School, Boston, MA 02115; [e]Clinomics Inc., 44919 Ulsan, Republic of Korea; [f]Laboratory of Aquatic Biomedicine, College of Veterinary Medicine, Seoul National University, 08826 Seoul, Republic of Korea; [g]Research Institute for Veterinary Science, College of Veterinary Medicine, Seoul National University, 08826 Seoul, Republic of Korea; [h]Hanwha Marine Biology Research Center, 63642 Jeju, Republic of Korea; [i]IT–Research Computing, Harvard Medical School, Boston, MA 02115; [j]Department of Mathematics, Harvard University, Cambridge, MA 02138; [k]Department of Zoology, University of Cambridge, CB2 3EJ Cambridge, United Kingdom; [l]National Institute of Biological Resources, 37242 Incheon, Republic of Korea; [m]College of Veterinary Medicine and Veterinary Medical Research Institute, Jeju National University, 63243 Jeju, Republic of Korea; and [n]Department of Chemistry and Chemical Biology, UNM Comprehensive Cancer Center, University of New Mexico, Albuquerque, NM 87131

Contributed by George M. Church, June 8, 2020 (sent for review December 24, 2019; reviewed by Manuel Corpas and Xiaohua Huang)

The endangered whale shark (*Rhincodon typus*) is the largest fish on Earth and a long-lived member of the ancient Elasmobranchii clade. To characterize the relationship between genome features and biological traits, we sequenced and assembled the genome of the whale shark and compared its genomic and physiological features to those of 83 animals and yeast. We examined the scaling relationships between body size, temperature, metabolic rates, and genomic features and found both general correlations across the animal kingdom and features specific to the whale shark genome. Among animals, increased lifespan is positively correlated to body size and metabolic rate. Several genomic traits also significantly correlated with body size, including intron and gene length. Our large-scale comparative genomic analysis uncovered general features of metazoan genome architecture: Guanine and cytosine (GC) content and codon adaptation index are negatively correlated, and neural connectivity genes are longer than average genes in most genomes. Focusing on the whale shark genome, we identified multiple features that significantly correlate with lifespan. Among these were very long gene length, due to introns being highly enriched in repetitive elements such as CR1-like long interspersed nuclear elements, and considerably longer neural genes of several types, including connectivity, activity, and neurodegeneration genes. The whale shark genome also has the second slowest evolutionary rate observed in vertebrates to date. Our comparative genomics approach uncovered multiple genetic features associated with body size, metabolic rate, and lifespan and showed that the whale shark is a promising model for studies of neural architecture and lifespan.

whale shark | lifespan | body size | metabolic rate | neural genes

The relationships between body mass, longevity, and basal metabolic rate (BMR) across diverse habitats and taxa have been researched extensively over the last century and have led to generalized rules and scaling relationships that explain many physiological and genetic trends observed across the tree of life. While the largest extant animals on the planet are aquatic, the impact of marine habitats on body size and other physiological and genetic characteristics is only beginning to be discovered (1). In an effort to better understand the selective pressures imposed on body size in marine environments, studies of endothermic aquatic mammals have shown that selection for larger body sizes has been driven by the minimization of heat loss (2). In ectothermic vertebrates, however, the relationship between environmental temperature and body size is more complex. In these species, metabolic rate is directly dependent on temperature, and decreased temperatures are correlated with decreased BMRs, decreased growth rates, longer generational times, and increased body sizes (3, 4).

The whale shark (*Rhincodon typus*) is the largest extant fish, reaches lengths of 20 m (5) and 42 tons in mass (6) and has a maximum lifespan estimated at 80 y (6). Worldwide populations have been declining, and the whale shark has been classified as an endangered species by the International Union for Conservation of Nature. Whale sharks are one of three species of filter-feeding sharks that use modified gill rakers to sieve plankton and small nektonic prey from the water column in a method

Significance

We sequenced and analyzed the genome of the endangered whale shark, the largest fish on Earth, and compared it to the genomes of 84 other species ranging from yeast to humans. We found strong scaling relationships between genomic and physiological features. We posit that these scaling relationships, some of which were remarkably general, mold the genome to integrate metabolic constraints pertaining to body size and ecological variables such as temperature and depth. Unexpectedly, we also found that the size of neural genes is strongly correlated with lifespan in most animals. In the whale shark, large gene size and large neural gene size strongly correlate with lifespan and body mass, suggesting longer gene lengths are linked to longer lifespans.

Author contributions: J.A.W., V.L., Y.S.C., J.B., and G.M.C. designed research; H.-M.K., S.W.K., W.H.H., Y.S.C., S.K., and J.-H.K. performed research; J.A.W., S.G.P., V.L., S.J., H.-M.K., Y.J., Y.B., J.H.J., S.L., A.K., and J.W.C. analyzed data; and J.A.W., S.G.P., V.L., S.J., A.M., J.S.E., J.B., and G.M.C. wrote the paper.

Reviewers: M.C., Cambridge Precision Medicine Limited; and X.H., University of California San Diego.

The authors declare no competing interest.

Data deposition: The whale shark whole-genome project data have been deposited at INSDC: International Nucleotide Sequence Database Collaboration (accession no. QPMN00000000). The version described in this paper is version QPMN01000000. DNA sequencing reads have been uploaded to the National Center for Biotechnology Information Sequence Read Archive (SRP155581). The C++ code used for the Markov Cluster (MCL) algorithm was uploaded to the GitHub repository (https://github.com/jsungwon/MCL-clustering).

[1]J.A.W., S.G.P., and V.L. contributed equally to this work.

[2]J.S.E., J.B., and G.M.C. contributed equally to this work.

[3]To whom correspondence may be addressed. Email: jsedwards@salud.unm.edu, jongbhak@genomics.org, or gchurch@genetics.med.harvard.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1922576117/-/DCSupplemental.

First published August 4, 2020.

Soft coral, *Dendronephthya gigantea,* genome assembly

GBE

The Draft Genome of an Octocoral, *Dendronephthya gigantea*

Yeonsu Jeon[1,2,†], Seung Gu Park[1,†], Nayun Lee[3,†], Jessica A. Weber[4,5], Hui-Su Kim[1], Sung-Jin Hwang[6], Seonock Woo[7], Hak-Min Kim[1,2], Youngjune Bhak[1,2], Sungwon Jeon[1,2], Nayoung Lee[3], Yejin Jo[3], Asta Blazyte[1], Taewoo Ryu[8], Yun Sung Cho[1,2,9], Hyunho Kim[10], Jung-Hyun Lee[7], Hyung-Soon Yim[7], Jong Bhak[1,2,9,10,*,‡], and Seungshic Yum[3,11,*,‡]

[1]Korean Genomics Industrialization and Commercialization Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea

[2]Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea

[3]Ecological Risk Research Division, Korea Institute of Ocean Science and Technology (KIOST), Geoje, Republic of Korea

[4]Department of Genetics, Harvard Medical School, Boston, Massachusetts

[5]Department of Biology, University of New Mexico

[6]Department of Life Science, Woosuk University, Republic of Korea

[7]Marine Biotechnology Research Center, Korea Institute of Ocean Science and Technology (KIOST), Busan, Republic of Korea

[8]APEC Climate Center, Busan, South Korea

[9]Clinomics Inc., Ulsan, Republic of Korea

[10]Personal Genomics Institute, Genome Research Foundation, Cheongju, Republic of Korea

[11]Faculty of Marine Environmental Science, University of Science and Technology (UST), Geoje, Republic of Korea

[†]These authors contributed equally to this work.

[‡]These authors jointly supervised this work.

*Corresponding authors: E-mails: jongbhak@genomics.org; syum@kiost.ac.kr.

Accepted: February 26, 2019

Data deposition: The octocoral whole genome and transcriptome project has been deposited at DDBJ/ENA/GenBank under the accession PRJNA507923 and PRJNA507943. DNA and RNA sequencing reads have been uploaded to the NCBI Read Archive under the accession (SRR8293699, and SRR8293698 and SRR8293935, and SRR8293936), respectively. The genome assembly has been deposited at DDBJ/ENA/GenBank under the accession RSEI01000000.

Abstract

Coral reefs composed of stony corals are threatened by global marine environmental changes. However, soft coral communities of octocorallian species, appear more resilient. The genomes of several cnidarians species have been published, including from stony corals, sea anemones, and hydra. To fill the phylogenetic gap for octocoral species of cnidarians, we sequenced the octocoral, *Dendronephthya gigantea*, a nonsymbiotic soft coral, commonly known as the carnation coral. The *D. gigantea* genome size is ~276 Mb. A high-quality genome assembly was constructed from PacBio long reads (29.85 Gb with 108× coverage) and Illumina short paired-end reads (35.54 Gb with 128× coverage) resulting in the highest N50 value (1.4 Mb) reported thus far among cnidarian genomes. About 12% of the genome is repetitive elements and contained 28,879 predicted protein-coding genes. This gene set is composed of 94% complete BUSCO ortholog benchmark genes, which is the second highest value among the cnidarians, indicating high quality. Based on molecular phylogenetic analysis, octocoral and hexacoral divergence times were estimated at 544 MYA. There is a clear difference in *Hox* gene composition between these species: unlike hexacorals, the Antp superclass Evx gene was absent in *D. gigantea*. Here, we present the first genome assembly of a nonsymbiotic octocoral, *D. gigantea* to aid in the comparative genomic analysis of cnidarians, including stony and soft corals, both symbiotic and nonsymbiotic. The *D. gigantea* genome may also provide clues to mechanisms of differential coping between the soft and stony corals in response to scenarios of global warming.

Key words: soft coral, genome, octocoral, nonsymbiotic coral, cnidarian, *Dendronephthya gigantea*.

Seven short-read sequencing platform comparison using KOREF

OXFORD (GIGA)n SCIENCE

GigaScience, 10, 2021, 1–9

doi: 10.1093/gigascience/giab014
Data Note

DATA NOTE

Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing

Hak-Min Kim[1], Sungwon Jeon[2,3], Oksung Chung[1], Je Hoon Jun[1], Hui-Su Kim[2], Asta Blazyte[2,3], Hwang-Yeol Lee[1], Youngseok Yu[1], Yun Sung Cho[1], Dan M. Bolser[4,*] and Jong Bhak[1,2,3,4,5,*]

[1]Clinomics Inc., Ulsan National Institute of Science and Technology (UNIST), UNIST-gil 50, Eonyang-eup, Ulju-gun, Ulsan, 44919, Republic of Korea; [2]Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), UNIST-gil 50, Eonyang-eup, Ulju-gun, Ulsan, 44919, Republic of Korea; [3]Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), UNIST-gil 50, Eonyang-eup, Ulju-gun, Ulsan, 44919, Republic of Korea; [4]Geromics Ltd., 222 Mill Road, Cambridge, CB1 3NF, United Kingdom and [5]Personal Genomics Institute (PGI), Genome Research Foundation, Osong saengmyong1ro, Cheongju, 28160, Republic of Korea

*Correspondence address: Dan M. Bolser, Geromics Ltd., 222 Mill Road, Cambridge, CB1 3NF, United Kingdom. E-mail: dan@geromics.co.uk http://orcid.org/0000-0002-3991-0859; Jong Bhak, Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), UNIST-gil 50, Eonyang-eup, Ulju-gun, Ulsan, 44919, Republic of Korea. Tel: +82-52-217-5329; E-mail: jongbhak@genomics.org http://orcid.org/0000-0002-4228-1299

Abstract

Background: DNBSEQ-T7 is a new whole-genome sequencer developed by Complete Genomics and MGI using DNA nanoball and combinatorial probe anchor synthesis technologies to generate short reads at a very large scale—up to 60 human genomes per day. However, it has not been objectively and systematically compared against Illumina short-read sequencers. **Findings:** By using the same KOREF sample, the Korean Reference Genome, we have compared 7 sequencing platforms including BGISEQ-500, DNBSEQ-T7, HiSeq2000, HiSeq2500, HiSeq4000, HiSeqX10, and NovaSeq6000. We measured sequencing quality by comparing sequencing statistics (base quality, duplication rate, and random error rate), mapping statistics (mapping rate, depth distribution, and percent GC coverage), and variant statistics (transition/transversion ratio, dbSNP annotation rate, and concordance rate with single-nucleotide polymorphism [SNP] genotyping chip) across the 7 sequencing platforms. We found that MGI platforms showed a higher concordance rate for SNP genotyping than HiSeq2000 and HiSeq4000. The similarity matrix of variant calls confirmed that the 2 MGI platforms have the most similar characteristics to the HiSeq2500 platform. **Conclusions:** Overall, MGI and Illumina sequencing platforms showed comparable levels of sequencing quality, uniformity of coverage, percent GC coverage, and variant accuracy; thus we

Received: 12 March 2020; **Revised:** 3 September 2020; **Accepted:** 16 February 2021

Molecules and Cells

Whole Genome Analysis of the Red-Crowned Crane Provides Insight into Avian Longevity

HyeJin Lee[1,8], Jungeun Kim[1,8], Jessica A. Weber[2,8], Oksung Chung[3], Yun Sung Cho[3], Sungwoong Jho[1], JeHoon Jun[3], Hak-Min Kim[4,5], Jeongheui Lim[6], Jae-Pil Choi[1], Sungwon Jeon[4,5], Asta Blazyte[4,5], Jeremy S. Edwards[7,*], Woon Kee Paek[6,*], and Jong Bhak[1,3,4,5,*]

[1]Personal Genomics Institute, Genome Research Foundation, Cheongju 28160, Korea, [2]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA, [3]Clinomics, Ulsan 44919, Korea, [4]KOGIC, Ulsan National Institute of Science and Technology, Ulsan 44919, Korea, [5]Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Korea, [6]National Science Museum, Ministry of Science and ICT, Daejeon 34143, Korea, [7]Chemistry and Chemical Biology, UNM Comprehensive Cancer Center, University of New Mexico, Albuquerque, NM 87131, USA, [8]These authors contributed equally to this work.
*Correspondence: jsedwards@salud.unm.edu (JSE); paekwk@naver.com (WKP); jongbhak@genomics.org (JB)
https://doi.org/10.14348/molcells.2019.0190
www.molcells.org

The red-crowned crane (*Grus japonensis*) is an endangered, large-bodied crane native to East Asia. It is a traditional symbol of longevity and its long lifespan has been confirmed both in captivity and in the wild. Lifespan in birds is known to be positively correlated with body size and negatively correlated with metabolic rate, though the genetic mechanisms for the red-crowned crane's long lifespan have not previously been investigated. Using whole genome sequencing and comparative evolutionary analyses against the grey-crowned crane and other avian genomes, including the long-lived common ostrich, we identified red-crowned crane candidate genes with known associations with longevity. Among these are positively selected genes in metabolism and immunity pathways (*NDUFA5, NDUFA8, NUDT12, SOD3, CTH, RPA1, PHAX, HNMT, HS2ST1, PPCDC, PSTK CD8B, GP9, IL-9R,* and *PTPRC*). Our analyses provide genetic evidence for low metabolic rate and longevity, accompanied by possible convergent adaptation signatures among distantly related large and long-lived birds. Finally, we identified low genetic diversity in the red-crowned crane, consistent with its listing as an endangered species, and this genome should provide a useful genetic resource for future conservation studies of this rare and iconic species.

Keywords: genome, longevity, red-crowned crane

INTRODUCTION

The red-crowned crane (*Grus japonensis*) is one of the rarest cranes in the world and is a symbol of longevity in East Asia. It has a maximum lifespan of 30 years in the wild and 65 years in captivity (Rasmussen and Engstrom, 2004), which is substantially longer than the average lifespan recorded in 144 other avian species (Supplementary Table S1). It is also one of the largest avian species (Chen et al., 2012), measuring on average 150-158 cm long with a 220-250 cm wingspan (del Hoyo et al., 1996), and weighing on average 8.9 kg (John and Dunning, 2008) (ranging from 4.8 to 10.5 kg). Since avian body sizes are known to be positively correlated to lifespan and negatively correlated to metabolic rate (McKechnie and Wolf, 2004; Scholander et al., 1950), it has previously been suggested that the longevity of the red-crowned crane is related to its low metabolic rate (Speakman, 2005).

Received 26 August, 2019; revised 31 October, 2019; accepted 18 December, 2019; published online 14 January, 2020

eISSN: 0219-1032

Bat, *Myotis rufoniger*, genome

PLOS ONE

RESEARCH ARTICLE

Myotis rufoniger genome sequence and analyses: *M. rufoniger's* genomic feature and the decreasing effective population size of *Myotis* bats

Youngjune Bhak[1,2], Yeonsu Jeon[1,2], Sungwon Jeon[1,2], Oksung Chung[3,4], Sungwoong Jho[3], JeHoon Jun[3,4], Hak-Min Kim[1,2], Yongsoo Cho[1,2], Changhan Yoon[1,5], Seungwoo Lee[6], Jung-Hoon Kang[7], Jong-Deock Lim[7], Junghwa An[8], Yun Sung Cho[1,2,3]*, Doug-Young Ryu[6]*, Jong Bhak[1,2,3,4]*

1 The Genomics Institute, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea, **2** Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea, **3** Personal Genomics Institute, Genome Research Foundation, Cheongju, Republic of Korea, **4** Geromics, Ulsan, Republic of Korea, **5** Department of Biomedical Science, School of Nano-Bioscience & chemical Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea, **6** BK21 PLUS Program for Creative Veterinary Science Research, Research Institute for Veterinary Science, and College of Veterinary Medicine, Seoul National University, Seoul, Republic of Korea, **7** National Research Institute of Cultural Heritage, Cultural Heritage Administration, Daejeon, Republic of Korea, **8** Animal Resources Division, National Institute of Biological Resources, Incheon, Republic of Korea

* joys0406@gmail.com (YSC); dyryu@snu.ac.kr (DYR); jongbhak@genomics.org (JB)

Citation: Bhak Y, Jeon Y, Jeon S, Chung O, Jho S, Jun J, et al. (2017) *Myotis rufoniger* genome sequence and analyses: *M. rufoniger's* genomic feature and the decreasing effective population size of *Myotis* bats. PLoS ONE 12(7): e0180418. https://doi.org/10.1371/journal.pone.0180418

Editor: Chongle Pan, Oak Ridge National Laboratory, UNITED STATES

Received: February 20, 2017

Accepted: May 23, 2017

Published: July 5, 2017

Data Availability Statement: All sequencing files are available from the National Center for Biotechnology Information (NCBI) database (566bp: SRX2755014, 574bp: SRX2755088).

Funding: This work was supported by 'the bioinformatics marker discovery analysis system using genomic big data' Research Fund (1.150014.01) of Ulsan National Institute of Science & Technology (UNIST). It was also supported by 'Software Convergence Technology Development Program' through the Ministry of

Abstract

Myotis rufoniger is a vesper bat in the genus *Myotis*. Here we report the whole genome sequence and analyses of the *M. rufoniger*. We generated 124 Gb of short-read DNA sequences with an estimated genome size of 1.88 Gb at a sequencing depth of 66× fold. The sequences were aligned to *M. brandtii* bat reference genome at a mapping rate of 96.50% covering 95.71% coding sequence region at 10× coverage. The divergence time of *Myotis* bat family is estimated to be 11.5 million years, and the divergence time between *M. rufoniger* and its closest species *M. davidii* is estimated to be 10.4 million years. We found 1,239 function-altering *M. rufoniger* specific amino acid sequences from 929 genes compared to other *Myotis* bat and mammalian genomes. The functional enrichment test of the 929 genes detected amino acid changes in melanin associated *DCT*, *SLC45A2*, *TYRP1*, and *OCA2* genes possibly responsible for the *M. rufoniger*'s red fur color and a general coloration in *Myotis*. *N6AMT1* gene, associated with arsenic resistance, showed a high degree of function alteration in *M. rufoniger*. We further confirmed that the *M. rufoniger* also has bat-specific sequences within *FSHB*, *GHR*, *IGF1R*, *TP53*, *MDM2*, *SLC45A2*, *RGS7BP*, *RHO*, *OPN1SW*, and *CNGB3* genes that have already been published to be related to bat's reproduction, lifespan, flight, low vision, and echolocation. Additionally, our demographic history analysis found that the effective population size of *Myotis* clade has been consistently decreasing since ~30k years ago. *M. rufoniger*'s effective population size was the lowest in *Myotis* bats, confirming its relatively low genetic diversity.

The giant Nomura's jellyfish genome assembly

Kim *et al. BMC Biology* (2019) 17:28
https://doi.org/10.1186/s12915-019-0643-7

BMC Biology

RESEARCH ARTICLE Open Access

The genome of the giant Nomura's jellyfish sheds light on the early evolution of active predation

Hak-Min Kim[1,2†], Jessica A. Weber[3,4†], Nayoung Lee[5†], Seung Gu Park[1], Yun Sung Cho[1,2,6], Youngjune Bhak[1,2], Nayun Lee[5], Yeonsu Jeon[1,2], Sungwon Jeon[1,2], Victor Luria[7], Amir Karger[8], Marc W. Kirschner[7], Ye Jin Jo[5], Seonock Woo[9,10], Kyoungsoon Shin[11], Oksung Chung[6,12], Jae-Chun Ryu[13], Hyung-Soon Yim[10], Jung-Hyun Lee[10], Jeremy S. Edwards[14], Andrea Manica[15], Jong Bhak[1,2,6,12*] and Seungshic Yum[5,9*]

Abstract

Background: Unique among cnidarians, jellyfish have remarkable morphological and biochemical innovations that allow them to actively hunt in the water column and were some of the first animals to become free-swimming. The class Scyphozoa, or true jellyfish, are characterized by a predominant medusa life-stage consisting of a bell and venomous tentacles used for hunting and defense, as well as using pulsed jet propulsion for mobility. Here, we present the genome of the giant Nomura's jellyfish (*Nemopilema nomurai*) to understand the genetic basis of these key innovations.

Results: We sequenced the genome and transcriptomes of the bell and tentacles of the giant Nomura's jellyfish as well as transcriptomes across tissues and developmental stages of the *Sanderia malayensis* jellyfish. Analyses of the *Nemopilema* and other cnidarian genomes revealed adaptations associated with swimming, marked by codon bias in muscle contraction and expansion of neurotransmitter genes, along with expanded Myosin type II family and venom domains, possibly contributing to jellyfish mobility and active predation. We also identified gene family expansions of *Wnt* and posterior *Hox* genes and discovered the important role of retinoic acid signaling in this ancient lineage of metazoans, which together may be related to the unique jellyfish body plan (medusa formation).

Conclusions: Taken together, the *Nemopilema* jellyfish genome and transcriptomes genetically confirm their unique morphological and physiological traits, which may have contributed to the success of jellyfish as early multi-cellular predators.

Keywords: Jellyfish mobility, Medusa structure formation, Scyphozoa, de novo genome assembly

Background

Cnidarians, including jellyfish and their predominantly sessile relatives the coral, sea anemone, and hydra, first appeared in the Precambrian Era and are now key members of aquatic ecosystems worldwide (Fig. 1a) [1]. Between 500 and 700 million years ago, jellyfish developed novel physiological traits that allowed them to become one of the first free-swimming predators. The life cycle of the jellyfish includes a small polypoid, sessile stage which reproduces asexually to form the mobile medusa form that can reproduce both sexually and asexually (Fig. 1c) [2]. The class Scyphozoa, or true jellyfish, are characterized by a predominant medusa life-stage consisting of a bell and venomous tentacles used for hunting and defense [3]. Jellyfish medusae feature a radially symmetric body structure, powered by readily identifiable cell types such as motor neurons and striated muscles that expand and contract to create the most energy-efficient swimming method in the animal kingdom [4, 5]. Over 95% water, jellyfish are osmoconformers that use ion gradients to deliver solutes to cells

* Correspondence: jongbhak@gmail.com; syum@kiost.ac.kr
†Hak-Min Kim, Jessica A. Weber and Nayoung Lee contributed equally to this work.
[1]Korean Genomics Industrialization Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea
[5]Ecological Risk Research Division, Korea Institute of Ocean Science and Technology (KIOST), Geoje 53201, Republic of Korea
Full list of author information is available at the end of the article

The leopard genome assembly

Kim *et al. Genome Biology* (2016) 17:211
DOI 10.1186/s13059-016-1071-4

Genome Biology

RESEARCH Open Access

Comparison of carnivore, omnivore, and herbivore mammalian genomes with a new leopard assembly

Soonok Kim[1†], Yun Sung Cho[2,3,4†], Hak-Min Kim[2,3†], Oksung Chung[4], Hyunho Kim[5], Sungwoong Jho[4], Hong Seomun[6], Jeongho Kim[7], Woo Young Bang[1], Changmu Kim[1], Junghwa An[6], Chang Hwan Bae[1], Youngjune Bhak[2], Sungwon Jeon[2,3], Hyejun Yoon[2,3], Yumi Kim[2], JeHoon Jun[4,5], HyeJin Lee[4,5], Suan Cho[4,5], Olga Uphyrkina[8], Aleksey Kostyria[8], John Goodrich[9], Dale Miquelle[10,11], Melody Roelke[12], John Lewis[13], Andrey Yurchenko[14], Anton Bankevich[15], Juok Cho[16], Semin Lee[2,3,17], Jeremy S. Edwards[18], Jessica A. Weber[19], Jo Cook[20], Sangsoo Kim[21], Hang Lee[22], Andrea Manica[23], Ilbeum Lee[24], Stephen J. O'Brien[14,25*], Jong Bhak[2,3,4,5*] and Joo-Hong Yeo[1*]

Abstract

Background: There are three main dietary groups in mammals: carnivores, omnivores, and herbivores. Currently, there is limited comparative genomics insight into the evolution of dietary specializations in mammals. Due to recent advances in sequencing technologies, we were able to perform in-depth whole genome analyses of representatives of these three dietary groups.

Results: We investigated the evolution of carnivory by comparing 18 representative genomes from across Mammalia with carnivorous, omnivorous, and herbivorous dietary specializations, focusing on Felidae (domestic cat, tiger, lion, cheetah, and leopard), Hominidae, and Bovidae genomes. We generated a new high-quality leopard genome assembly, as well as two wild Amur leopard whole genomes. In addition to a clear contraction in gene families for starch and sucrose metabolism, the carnivore genomes showed evidence of shared evolutionary adaptations in genes associated with diet, muscle strength, agility, and other traits responsible for successful hunting and meat consumption. Additionally, an analysis of highly conserved regions at the family level revealed molecular signatures of dietary adaptation in each of Felidae, Hominidae, and Bovidae. However, unlike carnivores, omnivores and herbivores showed fewer shared adaptive signatures, indicating that carnivores are under strong selective pressure related to diet. Finally, felids showed recent reductions in genetic diversity associated with decreased population sizes, which may be due to the inflexible nature of their strict diet, highlighting their vulnerability and critical conservation status.

Conclusions: Our study provides a large-scale family level comparative genomic analysis to address genomic changes associated with dietary specialization. Our genomic analyses also provide useful resources for diet-related genetic and health research.

Keywords: Carnivorous diet, Evolutionary adaptation, Leopard, Felidae, *De novo* assembly, Comparative genomics

* Correspondence: lgdchief@gmail.com; jongbhak@genomics.org; y1208@korea.kr
†Equal contributors
[14]Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg 199004, Russia
[2]The Genomics Institute, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea
[1]Biological and Genetic Resources Assessment Division, National Institute of Biological Resources, Incheon 22689, Republic of Korea
Full list of author information is available at the end of the article

Depression and suicide risk prediction using multi-omics

Bhak et al. *Translational Psychiatry* (2019)9:262
https://doi.org/10.1038/s41398-019-0595-2

Translational Psychiatry

ARTICLE Open Access

Depression and suicide risk prediction models using blood-derived multi-omics data

Youngjune Bhak[1,2,3], Hyoung-oh Jeong[1,2], Yun Sung Cho[3], Sungwon Jeon[1,2], Juok Cho[1,2], Jeong-An Gim[4], Yeonsu Jeon[1,2], Asta Blazyte[1], Seung Gu Park[1], Hak-Min Kim[1,2,3], Eun-Seok Shin[5], Jong-Woo Paik[6], Hae-Woo Lee[7], Wooyoung Kang[8], Aram Kim[8], Yumi Kim[3], Byung Chul Kim[3], Byung-Joo Ham[8,9,10], Jong Bhak[1,2,3,11] and Semin Lee[1,2]

Abstract

More than 300 million people worldwide experience depression; annually, ~800,000 people die by suicide. Unfortunately, conventional interview-based diagnosis is insufficient to accurately predict a psychiatric status. We developed machine learning models to predict depression and suicide risk using blood methylome and transcriptome data from 56 suicide attempters (SAs), 39 patients with major depressive disorder (MDD), and 87 healthy controls. Our random forest classifiers showed accuracies of 92.6% in distinguishing SAs from MDD patients, 87.3% in distinguishing MDD patients from controls, and 86.7% in distinguishing SAs from controls. We also developed regression models for predicting psychiatric scales with R^2 values of 0.961 and 0.943 for Hamilton Rating Scale for Depression–17 and Scale for Suicide Ideation, respectively. Multi-omics data were used to construct psychiatric status prediction models for improved mental health treatment.

Introduction

Suicide and depression are major health hazards, resulting in the death of one person every 40 s globally[1,2]. They are complex and intertwined phenomena: ~4% of individuals diagnosed with depression commit suicide, and more than half of the persons who attempt suicide meet the criteria of depression[3]. The suicide rate in South Korea (25.8 deaths per 100,000 persons) is among the highest worldwide and is 2.30 times higher than the average of the Organization for Economic Co-operation and Development (OECD) countries (11.2 deaths per 100,000 persons). South Korea has been ranked second among the OECD countries in terms of suicide rates. Notably, the suicide rate for women in South Korea is the highest (14.7 deaths per 100,000 women) among the OECD countries (average 4.86 deaths per 100,000 women)[4]. Hence, predicting depression and suicide risk is a global problem, with exceptional importance in South Korea. Therefore, developing effective models for predicting depression and suicidality may elucidate breakthrough treatments.

The current depression and suicide prediction methods rely on self-reported measures such as questionnaires and interviews, which can be too subjective; and people with depression and suicidal ideation may not be honest about expressing their thoughts[5]. Thus, health records or neural representations have been adopted, with machine learning techniques, to predict the risk of depression and suicide[6,7]. Identifying highly accurate biomarkers would also be an ideal solution that would give an insight to our understanding of depression and suicide. Since the brain is the target organ in psychiatry, brain-based biomarkers have

Correspondence: Jong Bhak (jongbhak@gmail.com) or Semin Lee (seminlee@unist.ac.kr)
[1]Korean Genomics Industrialization and Commercialization Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea
[2]Department of Biomedical Engineering, School of Life Sciences, UNIST, Ulsan 44919, Republic of Korea
Full list of author information is available at the end of the article.
These authors contributed equally: Youngjune Bhak, Hyoung-oh Jeong

Human Genetics (2020) 139:557–568
https://doi.org/10.1007/s00439-020-02132-8

ORIGINAL INVESTIGATION

Decoding a highly mixed Kazakh genome

Madina Seidualy[1] · Asta Blazyte[1] · Sungwon Jeon[1,2] · Youngjune Bhak[1,2] · Yeonsu Jeon[1,2] · Jungeun Kim[3] · Anders Eriksson[4,5] · Dan Bolser[6] · Changhan Yoon[1,2] · Andrea Manica[7] · Semin Lee[1,2] · Jong Bhak[1,2,3,8]

Received: 18 December 2019 / Accepted: 5 February 2020 / Published online: 19 February 2020

Abstract
We provide a Kazakh whole genome sequence (MJS) and analyses with the largest comparative Kazakh genomic data available to date. We found 102,240 novel SNVs and a high level of heterozygosity. ADMIXTURE analysis confirmed a significant proportion of variations in this individual coming from all continents except Africa and Oceania. A principal component analysis showed neighboring Kalmyk, Uzbek, and Kyrgyz populations to have the strongest resemblance to the MJS genome which reflects fairly recent Kazakh history. MJS's mitochondrial haplogroup, J1c2, probably represents an early European and Near Eastern influence to Central Asia. This was also supported by the heterozygous SNPs associated with European phenotypic features and strikingly similar Kazakh ancestral composition inferred by ADMIXTURE. Admixture (*f3*) analysis showed that MJS's genomic signature is best described as a cross between the Neolithic East Asian (Devil's Gate1) and the Bronze Age European (Halberstadt_LBA1) components rather than a contemporary admixture.

Introduction

Recently, a wide variety of genome sequencing technologies have become available heralding a new era of personal genomics (Lander et al. 2001) and many large population genome projects have been carried out. These include the 1000 Genomes Project (Abecasis et al. 2010), the UK's 10,000 and 100,000 Genomes Projects (Walter et al. 2015; Samuel and Farsides 2017), the Genome of the Netherlands (Boomsma et al. 2014), the Estonian Biocentre's Human Genome Diversity Panel (EGDP) (Pagani et al. 2016), the Simons Genome Diversity Project (SGDP) (Mallick et al. 2016), the Genome Russia project (Oleksyk et al. 2015), 1070 Japanese genomes (Nagasaki et al. 2015), and the Korean Reference (KOREF) and variome projects (Cho et al. 2016; Kim et al. 2018). The Personal Genome Project (PGP) (Ball et al. 2012) is perhaps the largest genome project in terms of openness and inclusiveness and aims to map all personal and ethnic genomes. However, there remain many practical issues for mapping and accurately analyzing all ethnic groups worldwide. One problem is suitable representation of highly admixed genomes (Medina-Gomez et al. 2015; Guryev 2017). Although the 1000 Genomes Project database has been expanding by adding more ethnic representatives, it currently contains only 2504 individuals from 26 populations (phase 3) and lacks much ethnic diversity including an absence of genomes from Central Asian populations (Sudmant et al. 2015). Other initiatives such as the SGDP and the EGDP include only a small number of Central Asian population representatives (Pagani et al. 2016; Mallick et al. 2016). Central Asian populations can be good targets for adding highly admixed samples to our knowledge base of the major and relatively homogeneous ethnic groups. Among many Central Asian countries, Kazakhstan is at the border of ethnically European and Asian nations (Mostafa 2013). Therefore, demographic inference from Kazakh whole genomes is of special value. We can use Kazakh genomic data as an independent line of evidence that complements inference from archeological and written histories, to understand the roots of the diverse phenotypic features and relationships with other

Madina Seidualy, Asta Blazyte and Sungwon Jeon have contributed equally to this work.

Sequence data from this article have been deposited to NCBI SRA database under accession No. SRS2904218 and NCBI BioSample database under accession No. SAMN08442411.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00439-020-02132-8) contains supplementary material, which is available to authorized users.

✉ Jong Bhak
jongbhak@genomics.org

Extended author information available on the last page of the article

Iranian ancient genome

SCIENTIFIC REPORTS

OPEN

The genetics of an early Neolithic pastoralist from the Zagros, Iran

M. Gallego-Llorente[1], S. Connell[2], E. R. Jones[1], D. C. Merrett[3], Y. Jeon[4,5], A. Eriksson[1,6], V. Siska[1], C. Gamba[2,7], C. Meiklejohn[8], R. Beyer[9], S. Jeon[4,5], Y. S. Cho[4,5], M. Hofreiter[10], J. Bhak[4], A. Manica[1,*] & R. Pinhasi[2,*]

Received: 28 May 2016
Accepted: 15 July 2016
Published: 09 August 2016

The agricultural transition profoundly changed human societies. We sequenced and analysed the first genome (1.39x) of an early Neolithic woman from Ganj Dareh, in the Zagros Mountains of Iran, a site with early evidence for an economy based on goat herding, ca. 10,000 BP. We show that Western Iran was inhabited by a population genetically most similar to hunter-gatherers from the Caucasus, but distinct from the Neolithic Anatolian people who later brought food production into Europe. The inhabitants of Ganj Dareh made little direct genetic contribution to modern European populations, suggesting those of the Central Zagros were somewhat isolated from other populations of the Fertile Crescent. Runs of homozygosity are of a similar length to those from Neolithic farmers, and shorter than those of Caucasus and Western Hunter-Gatherers, suggesting that the inhabitants of Ganj Dareh did not undergo the large population bottleneck suffered by their northern neighbours. While some degree of cultural diffusion between Anatolia, Western Iran and other neighbouring regions is possible, the genetic dissimilarity between early Anatolian farmers and the inhabitants of Ganj Dareh supports a model in which Neolithic societies in these areas were distinct.

The agricultural transition started in a region comprising the Ancient Near East and Anatolia ~12,000 years ago with the first Pre-Pottery Neolithic villages and the first domestication of cereals and legumes[1,2]. Archaeological evidence suggests a complex scenario of multiple domestications in a number of areas[3], coupled with examples of trade[4]. Ancient DNA (aDNA) has revealed that this cultural package was later brought into Europe by dispersing farmers from Anatolia (so called 'demic' diffusion, as opposed to non-demic cultural diffusion[5,6]) ~8,400 years ago. However a lack of aDNA from early Neolithic individuals from the Near East leaves a key question unanswered: was the agricultural transition developed by one major population group spanning the Near East, including Anatolia and the Central Zagros Mountains; or was the region inhabited by genetically diverse populations, as is suggested by the heterogeneous mode and timing of the appearance of early domesticates at different localities?

To answer this question, we sequenced the genome of an early Neolithic female from Ganj Dareh, GD13a, from the Central Zagros (Western Iran), dated to 10000-9700 cal BP[7], a region located at the eastern edge of the Near East. Ganj Dareh is well known for providing the earliest evidence of herd management of goats beginning at 9,900 BP[7–9]. It is a classic mound site at an altitude of ~1400 m in the Gamas-Ab Valley of the High Zagros zone in Kermanshah Province, Western Iran. It was discovered in the 1960s during survey work and excavated over four seasons between 1967 and 1974. The mound, ~40 m in diameter, shows 7 to 8 m of early Neolithic cultural deposits. Five major levels were found, labelled A through E from top to bottom. Extended evidence showed a

[1]Department of Zoology, University of Cambridge, Cambridge, CB2 3EJ, UK. [2]School of Archaeology and Earth Institute, University College Dublin, Belfield, Dublin 4, Ireland. [3]Department of Archaeology, Simon Fraser University, Burnaby, BC V5A 1S6, Canada. [4]The Genomics Institute, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea. [5]Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea. [6]Integrative Systems Biology Laboratory, Division of Biological and Environmental Sciences & Engineering, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia. [7]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, Copenhagen 1350, Denmark. [8]Department of Anthropology, University of Winnipeg, Winnipeg, MB R3B 2E9, Canada. [9]McDonald Institute for Archaeological Research, University of Cambridge, Cambridge CB2 3ER, UK. [10]Evolutionary Adaptive Genomics, Institute for Biochemistry and Biology, Department of Mathematics and Natural Sciences, University of Potsdam, Karl-Liebknechtstraße 24-25, Potsdam, 14476, Germany. *These authors jointly supervised this work. Correspondence and requests for materials should be addressed to M.G.-L. (email: mg632@cam.ac.uk) or A.M. (email: am315@cam.ac.uk) or R.P. (email: ron.pinhasi@ucd.ie)